

As desigualdades regionais de renda entre indivíduos no Brasil

Área de interesse: Área 1 – Economia Regional.

Alexandre Rands Coelho Barros

Ph.D em Economia

Universidade Federal de Pernambuco

University of Illinois, em Urbana-Champaign-EUA, Ph.D. em Economia, 1991.

University of Illinois, em Urbana-Champaign-EUA, mestrado em Economia, 1986.

Universidade Federal de Pernambuco, graduado em Economia, 1984.

Rua do Bom Jesus, 220 – Bairro do Recife - Recife – PE

Cep.: 50.030-170

[alexandre.rands@datametrica.com.br](mailto:alexandre.rands@datametrica.com.br)

Fone: (81) 3316.2613 Fax: (81) 3316.2632

Recife-Brazil

Abril 2013

### Abstract

This paper argues that the existing empirical evidence that there are inequalities in individual labor incomes in Brazil as a consequence of differences in regional residence is spurious. Such results arise from the correlation of education with omitted variables and the existence of non-linearities on the relationship between schooling and income. When these specific sources of distortions are eliminated, such evidences disappear. The paper relies on a method that: (i) better specifies the adequate Mincer equation to be estimated and (ii) avoid generating spurious results because of poor treatment of some non-linearities and dataset compositions. It uses data from the 2000 Census to estimate expanded Mincer equations that avoid these distortions and relies on a simulation that builds on Oaxacca decomposition to generate a new estimated relative regional income. The empirical results indicate that the Brazilian labor market seems to arbitrate fairly among regions and as such that there is no regional inequalities emerging from differences in returns to labor with the similar features in different regions. These results strength the hypothesis that regional inequalities in Brazil arise from uneven distribution of human capital among regions, rather than any other factor of production or even economic discrimination.

Keywords: Regional inequality, labor income, Mincer equation.

### Resumo

Esse artigo argumenta que as evidências empíricas que tentam dar suporte à hipótese de que existem desigualdades regionais de rendas individuais no Brasil sofrem de distorções metodológicas que respondem pelos resultados encontrados. Um método alternativo para proceder com tais estimações é apresentado. A existência de não linearidades na relação entre educação e renda, e ponderações espúrias entre indivíduos com diferentes atributos são as distorções corrigidas. Uma Equação de Mincer expandida com a introdução dessas correções é utilizada em estimações com dados do Censo de 2000. Os resultados são utilizados em simulações que são construídas a partir de uma expansão da Decomposição de Oaxacca para comparar as rendas entre indivíduos residentes no Nordeste, por um lado, e no Sul ou Sudeste, pelo outro. Tais exercícios mostram que não há evidência em suporte da hipótese de que existem desigualdades de rendas individuais geradas apenas por decorrência do local de residência. A partir desses resultados fortalece-se a hipótese de que as desigualdades regionais é mais fortemente associada a distribuição desigual de atributos produtivos entre indivíduos do que por causa de diferenças entre distribuição regional de outros fatores de produção ou mesmo por causa do fenômeno econômico da discriminação.

Palavras chaves: Desigualdades regionais, rendimento do trabalho, equação de Mincer.

JEL: J31; R11.

---

## 1. Introdução

As desigualdades regionais no Brasil têm impressionado todas as pessoas que se dedicam ao seu estudo. Às vezes é difícil entender como em um mesmo país podem existir estados cuja renda per capita seja quase um quarto da de outros. Segundo dados do IBGE, em 2008 o PIB per capita do Piauí, unidade mais pobre da federação, representou apenas 11,7% do calculado para o Distrito Federal e 22% daquele de São Paulo, os dois estados mais ricos do país. Tais diferenças têm sido apontadas como evidências óbvias de que há grandes disparidades regionais no Brasil.

Além dos dados econômicos agregados como os apresentados acima, há várias outras estatísticas que têm servido para revelar a magnitude das desigualdades regionais brasileiras. Estatísticas sociais como percentual da população analfabeta ou vivendo abaixo da linha da pobreza e mesmo de residentes em domicílios classificados pelo IBGE como precários são algumas dessas estatísticas que mostram existir grandes diferenças no padrão de vida da população entre estados brasileiros.<sup>1</sup>

Várias têm sido as causas apontadas para esse fenômeno. As principais têm sido a existência de grandes disparidades no nível de capital humano entre as regiões.<sup>2</sup> Em estados ou municípios mais pobres as pessoas têm menos educação formal. Enquanto em São Paulo uma pessoa de 25 anos ou mais na média passou 7,94 anos na escola, no Piauí indivíduos nessa mesma faixa etária teria despendido apenas 4,99 anos dedicada ao estudo formal.<sup>3</sup> Ao observar essas estatísticas, uma primeira curiosidade que salta aos olhos é que as diferenças proporcionais entre anos de estudo é menor do que entre os PIBs per capita, mas ainda assim essa hipótese pode ser adequada, se as elasticidades tiverem valores específicos.

Diante dessa possibilidade de que haja disparidades regionais elevadas e que possam ser explicadas apenas por anos médios de estudo, alguns estudos têm sofisticado um pouco mais a análise desse fenômeno para verificar se quando se isola o efeito das diferenças em anos de estudo e outros atributos individuais simples, como experiência e distribuição das pessoas por raças e gêneros distintos, ainda assim há diferenças nas rendas entre elas por causa da região em que elas vivem. Esses estudos até então têm confirmado a existência de desigualdades regionais no Brasil no sentido de que um indivíduo que vive no Nordeste é mais pobre do que aqueles vivendo no Sul e Sudeste, mesmo que todos eles tenham os mesmos atributos individuais que poderiam afetar sua renda.<sup>4</sup>

Esse artigo visa exatamente mostrar que os estudos até então realizados cuja conclusão aponta para a existência de desigualdades regionais mesmo quando se elimina as diferenças em atributos produtivos individuais na verdade estão ficando com um resquício de desigualdade por causa de algumas fragilidades metodológicas. A utilização de um método que as supere pode não só chegar a conclusões distintas como serve também para revelar os fatores que têm gerado as desigualdades encontradas até então.

---

<sup>1</sup> Barros (2010 e 2011a) trazem algumas estatísticas recentes para outros indicadores de disparidades regionais.

<sup>2</sup> Ver, por exemplo, Barros (2011b).

<sup>3</sup> Dados para 2007 do IBGE, PNAD.

<sup>4</sup> Ver, por exemplo, Hoffmann (2011) e Barros, Carvalho, Franco e Mendonça (2010).

O artigo está organizado como segue. Na próxima seção uma regressão simples mostra como as desigualdades regionais parecem persistir mesmo após a inclusão de vários atributos que afetam as rendas individuais. Na seção seguinte, apresentam-se algumas críticas ao método utilizado na seção 2 e justifica-se a possibilidade de se encontrar disparidade gerada por local de residência, mesmo quando ela não existe. A seção 4 apresenta uma derivação mais adequada do modelo a se estimar e a seção 5 discute uma metodologia que deve corrigir pelo menos parcialmente os problemas encontrados, além de trazer novas estimações econométricas que geram conclusões diferentes das tradicionalmente obtidas. A seção 6 resume as principais conclusões do trabalho.

## 2. Estimções iniciais

Normalmente os estudos que identificam a existência de disparidades regionais de renda, mesmo quando se corrige para diferenças nos atributos individuais, recorrem a uma equação Minceriana para determinar a renda individual do trabalho. Numa versão padrão, mas não única, ela pode ser apresentada como:

$$\ln w_i = \alpha_0 + \alpha_1 S_i + \alpha_2 X_i + \alpha_3 X_i^2 + \alpha_4 R_i + \sum_{j=1}^n \beta_j Z_{ji} + e_i \quad (1)$$

Onde  $\ln w_i$  representa o logaritmo natural do rendimento do trabalho do indivíduo  $i$ ,  $S_i$  os anos de estudo também do indivíduo  $i$  e  $X_i$  a experiência desse mesmo indivíduo no mercado de trabalho.  $R_i$  representa uma *dummy* regional com 1 para os indivíduos que moram nas regiões nas quais eles não são vítimas do problema regional e 0 para os residentes nas regiões mais atrasadas. Normalmente o Sul e o Sudeste são as regiões no primeiro grupo e o Nordeste e o Norte as incluídas no segundo. As demais variáveis  $Z_j$  (para  $j$  pertencente ao conjunto dos números inteiros positivos e menor ou igual a  $n$ ) são outros atributos do indivíduo e  $e_i$  é um erro aleatório, que normalmente se toma como sendo independentemente distribuído com média zero e variância finita.

A variável que identifica a existência de desigualdade regional, mesmo quando se isola para atributos individuais é a *dummy*  $R$ . Se  $\alpha_4 > 0$  isso significa que há diferenças nas rendas individuais por causa da região de residência mesmo quando se isola para os demais fatores representados pelas variáveis incluídas na equação (1). Assim o teste da hipótese de que  $\alpha_4 = 0$  contra a hipótese de que esse coeficiente é maior do que zero tem sido a estratégia frequentemente utilizada para mostrar que existe disparidade regional que emerge não só dos atributos individuais normalmente identificados como determinantes das rendas individuais.

Os dados utilizados para estimar a equação (1) têm sido normalmente obtidos das PNADs, vários anos, ou censos populacionais, ambas as pesquisas realizadas pelo IBGE. Nessa seção utilizar-se-á dados do censo de 2000 para realizar tal estimação. Os dados do censo têm algumas vantagens sobre os das PNADs, pois eles permitem atrelar o indivíduo ao município em que reside. Essa característica será útil no decorrer desse trabalho. Os dados nesse trabalho incluem indivíduos apenas das regiões Sul, Sudeste e Nordeste. A *dummy* regional tem valor 1,0 para indivíduos residentes no Sul ou Sudeste. Indivíduos residentes nas cidades capitais do Rio de Janeiro e de São Paulo foram deixados de fora da amostra por motivos que serão discutidos melhor adiante.

O método de estimação tem variado também entre trabalhos. O mais simples é a aplicação de Mínimos Quadrados Ordinários diretamente aos dados por indivíduos, extraídos dos microdados das pesquisas mencionadas. Somente os indivíduos com rendimento do trabalho positivo são incluídos nessas estimações. Esse método, contudo, apresenta problema de viés de seleção, pois os indivíduos que por alguma razão receberiam uma renda baixa quando se compara a outros com os mesmos atributos podem optar por ficar de fora do mercado de trabalho e por tal não serem incluídos na amostra. Essa lógica no processo de seleção dos indivíduos que são incluídos na regressão viesam os resultados das estimações.<sup>5</sup> Assim, frequentemente se utiliza o método dos Mínimos Quadrados Ordinários com correção para o viés de seleção a partir de método sugerido por Heckman (1979). Esse será o método utilizado nas estimações a seguir, incluindo-se as rendas do não trabalho e uma *dummy* com um para chefes de domicílio como instrumentos no primeiro estágio, o PROBIT, que é utilizado para se encontrar a Inversa de Mills. Vale lembrar que mesmo nesse primeiro estágio, todas as demais variáveis dos MQO para o segundo estágio foram também incluídas.

Os resultados assim como uma melhor especificação das variáveis incluídas na equação (1) aparecem na tabela 1. Numa primeira versão incluíram-se nesse segundo estágio apenas as variáveis explicitamente inseridas na equação (1), sem nenhuma  $Z_j$  adicional. Numa segunda versão, que aparece na tabela 2, algumas variáveis  $Z_j$  adicionais foram incluídas, sendo elas comumente encontradas na literatura.

Tabela 1  
Resultados das estimações por Mínimos quadrados ordinários  
da versão mais simples da equação (1)

Variável	Coefficiente	Estatística T	Nível de Significância
Constante	-0,4916	-158,2646	0,0000
Anos de estudo	0,1059	82,1905	0,0000
Experiência	0,0791	462,7672	0,0000
Experiência ao quadrado	-0,0009	-295,9108	0,0000
<i>Dummy</i> regional (Sul e Sudeste=1)	0,4183	350,3913	0,0000
Inversa de Mills	0,6493	37,1129	0,0000
$R^2=$	0,37	Amostra=	3.703.455

Nota: As Estatísticas t e o nível de significância foram calculados a partir de correção para heteroscedasticidade pelo método de White (1980).

Fonte: Cálculo próprio.

Em ambas as estimações, a hipótese nula de que  $\alpha_4=0$  é rejeitada a níveis de significância elevados. Na estimação mais simples da tabela 1, esse coeficiente estimado é 0,4183. Na segunda estimação, com mais variáveis, ele cai para 0,35506. As estatísticas t-Student para a hipótese nula de que esses coeficientes são iguais a 0 são superiores a 300 em ambos os casos. Ou seja, a hipótese de que a região de residência dos indivíduos não determina suas rendas, mesmo após se isolar os principais efeitos que poderiam determinar as diferenças de renda do trabalho, é rejeitada nessas estimações. Segundo esse método, estatisticamente isso significa que a região de residência é um importante determinante da

<sup>5</sup> Esse problema foi apontado por Heckman (1979) e desde então desencadeou uma enorme literatura sobre o assunto.

renda e por tal as desigualdades regionais existiriam independentemente desses outros determinantes das rendas individuais.

Tabela 2  
Resultados das estimações por Mínimos quadrados ordinários  
da versão da equação (1) que inclui outras variáveis

Variável	Coefficiente	Estatística T	Nível de Significância
Constante	-0,83255	-191,6872	0,00000
Anos de estudo	0,09362	719,9414	0,00000
Experiência	0,07659	460,8750	0,00000
Experiência ao quadrado	-0,00094	-327,1294	0,00000
Dummy regional (Sul e Sudeste=1)	0,35506	324,0403	0,00000
Dummy sexo (masc=1)	0,36543	420,3068	0,00000
Logaritmo natural do tamanho da família	0,09248	144,1098	0,00000
Dummy branco = 1	0,17925	206,4995	0,00000
Dummy negro = 1	-0,02075	-14,3149	0,00000
Dummy amarelo = 1	0,43093	62,0814	0,00000
Dummy índio = 1	0,02627	3,5063	0,00045
Dummy conta própria =1	-0,02404	-21,5554	0,00000
Dummy empregador = 1	0,87814	286,3917	0,00000
Dummy sem carteira = 1	-0,25906	-246,4947	0,00000
Dummy funcionário público = 1	0,39372	186,2106	0,00000
Inversa de Mills	1,39075	92,8912	0,00000
R <sup>2</sup> =	0,46	Amostra=	3.703.455

Nota: As Estatísticas t e o nível de significância foram calculados a partir de correção para heteroscedasticidade pelo método de White (1980).

Fonte: Cálculo próprio.

Os valores estimados para os coeficientes também são elevados. Em ambas as regressões o exponencial desses valores multiplicados por um gera valores que representam cerca de 10% da média da variável rendimento do trabalho encontrada na amostra. Ou seja, o fato de alguém residir no Nordeste reduz em cerca de 10% a renda dele, segundo esses cálculos, considerando-se o indivíduo com a renda média das duas regiões.

### 3. Fonte de distorções nessas estimações

Há várias possíveis fontes de problemas na estimação da equação (1). Várias delas têm sido discutidas na literatura, principalmente no contexto de suas consequências para a estimação do retorno marginal da educação. As principais são: (i) a endogeneidade da escolaridade por causa da relação entre habilidades individuais natas e persistência na escola;<sup>6</sup>(ii) existência de não linearidades na relação entre escolaridade e rendimento do trabalho;<sup>7</sup> (iii) possível não separabilidade dos efeitos de escolaridade e experiência.<sup>8</sup> Entretanto, o desempenho dessas equações têm sido razoáveis, desde que algumas correções

<sup>6</sup> Essa fragilidade foi apontada inicialmente por Griliche (1977)

<sup>7</sup> Ver por exemplo Card (1999) e Lamieux (2006).

<sup>8</sup> Ver Belzil (2006) e Heckman, Lochman e Todd (2007).

sejam introduzidas,<sup>9</sup> devendo ser elas as adequadas para o problema analisado a partir de tal estimação.

No caso particular dos objetivos propostos nesse trabalho, dois problemas podem estar gerando os resultados regionais acima apresentados de forma espúria. Um primeiro e óbvio é que algumas variáveis importantes podem estar sendo omitidas no modelo e a *dummy* regional pode estar servindo de *proxy* para elas. A segunda é que pode haver não linearidades nas relações capturadas pela equação (1) que estejam sendo desprezadas e por tal podem ser as verdadeiras responsáveis pelos resultados obtidos. Essas duas fontes de distorções serão objeto de discussão nessa seção.

### 3.1. Não linearidades desprezadas

É possível que algumas das relações incluídas na equação (1) de fato não sejam lineares, como apresentadas. Elas podem ter segundas derivadas, ou até mesmo derivadas de maior ordem, diferentes de zero e com isso gerarem relações não lineares entre as variáveis independentes e a renda individual do trabalho. Quando essas relações são forçadas a serem lineares nas estimações, as *dummies* podem elevar a capacidade do método dos mínimos quadrados ordinários de capturar a variação da variável dependente, gerando distorções que não refletem o processo de geração de dados.

Isso vai ocorrer no caso particular da *dummy* regional se a média dos valores da variável independente, cuja relação com a dependente é não-linear, for diferente nas duas regiões que são identificadas na estimação. Nesse caso a *dummy* regional pode aparecer como significativamente diferente de zero, mesmo que ela não o seja no processo de geração de dados. Ou seja, ela vai espuriamente aparecer como relevante na explicação dos dados.

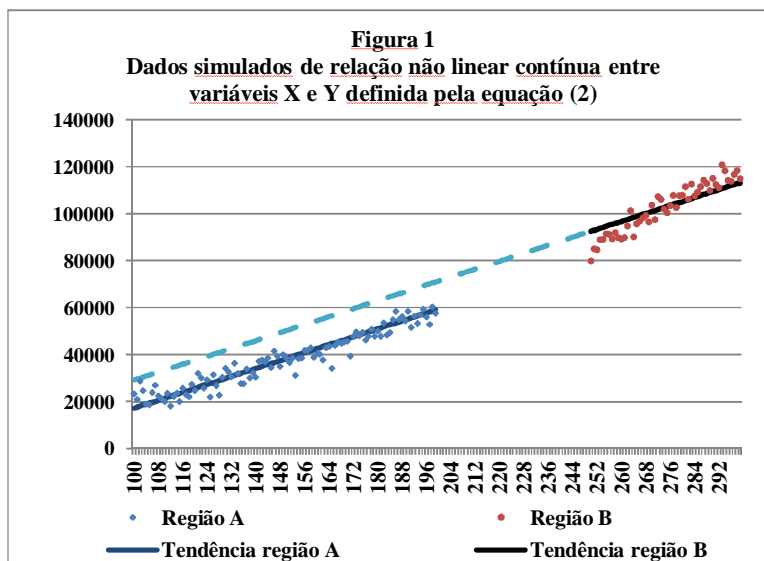
Um exemplo numérico pode mostrar concretamente esse problema. Suponha uma relação entre duas variáveis, X e Y, como a apresentada na figura 1. Nesse caso, por construção, a equação que define a relação entre essas variáveis é:

$$Y_i = 50 + 100X_i + X_i^2 + v_i \quad (2)$$

Onde  $v$  é uma variável aleatória independente e identicamente distribuída com distribuição Normal, média zero e desvio padrão igual a 3000. Ou seja,  $v \sim N(0,3000)$ . Supõe-se nesse caso, por simplificação na apresentação, que não há intersecção entre os dados para a variável X da região A e da região B. Na região A os valores para X estendem-se entre 100 e 200. Na região B eles estendem-se entre 250 e 300. Assim, os indivíduos nas duas regiões ficam bem apartados quando o eixo do X está representado de forma contínua entre 100 e 300, como na figura 1.

---

<sup>9</sup> Ver Lamieux (2006).



Nesse exemplo, sendo a função da equação (2) supostamente desconhecida, estimou-se outra relação linear definida como:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 R_i + v_i \quad (3)$$

Nessa equação linear incluiu-se uma *dummy* regional R, como feito acima com os dados do Censo de 2000. Os resultados estão apresentados na tabela 3. Como pode se ver, surgiu um impacto ou comportamento regional diferenciado em uma relação que por construção não detinha tal característica. Apenas a má especificação da equação a ser estimada gerou esse resultado espúrio.<sup>10</sup>

Esse exemplo mostra o que pode justificar o impacto e significância da *dummy* regional nas estimações apresentadas nas tabelas 1 e 2 da seção anterior. Assim, um pouco mais de aprofundamento da análise é necessário para verificar se realmente o local de residência do indivíduo tem efeito na sua renda.

Tabela3  
 Estimações por mínimos quadrados de dados simulados  
 para duas regiões a partir da equação (2)

Variável	Coefficiente	Estatística t	Significância
Constante	-25.154,89	-12,584	0,00000
X	423,02	32,292	0,00000
DR	11.935,33	6,710	0,00000

Fonte: Elaboração própria

<sup>10</sup> A figura 1 traz a tendência estimada com a suposição de quebra regional nessa tendência, mas mantendo-se o mesmo coeficiente  $\beta_1$  da equação (3). Apenas somou-se  $\beta_2$  na tendência da região B.



### 3.2. Variáveis excluídas

A exclusão de variáveis que possam ter alguma correlação com outras incluídas no modelo, mas sem relevância efetiva para explicar a determinação da dependente, pode gerar distorções nos resultados encontrados. No caso particular da *dummy* regional, se a variável excluída tiver médias diferentes nas duas regiões já seria suficiente para fazer com que o coeficiente estimado para a *dummy* regional fosse significativamente diferente de zero.

O exemplo seguinte mostra de forma mais formal esse fato. Suponha que uma variável  $Y$  seja determinada pela seguinte equação:

$$Y = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \delta_3 X_3 + e \quad (4)$$

Nesse caso  $X_1$ ,  $X_2$  e  $X_3$  são variáveis exógenas e  $e$  um termo aleatório, independente e identicamente distribuído, com média zero e variância finita, sendo  $E(X_i e) = 0$ , para  $i = 1; 2$  ou  $3$ . As letras gregas são coeficientes, todos diferentes de zero, com exceção de  $\delta_3$ , que, por construção, é igual a zero. Ou seja, não há causalidade da variável  $X_3$  para  $Y$ .

Suponha também que:

$$X_3 = \gamma_0 + \gamma_1 X_2 + v \quad (5)$$

Onde as letras gregas são coeficientes, todos diferentes de zero, e  $v$  é uma variável aleatória com média zero e variância finita. Se a equação estimada, em vez de ter a especificação da equação (4), alternativamente for definida como:

$$Y = \delta'_0 + \delta'_1 X_1 + \delta'_3 X_3 + e' \quad (4')$$

Então, nesse caso, teremos que:

$$\delta'_3 = \frac{\delta_2}{\gamma_1} + \delta_3 = \frac{\delta_2}{\gamma_1} \neq 0 \quad (6)$$

Ou seja, a variável  $X_3$ , a partir desse erro de especificação passa a ter coeficiente estimado que converge para valor diferente de zero por causa da omissão da variável  $X_2$  no modelo estimado. Se  $X_3$  for a *dummy* regional, então essa variável passará a ter efeito significativamente diferente de zero, como visto nas estimações empíricas da equação (1) apresentadas nas tabelas (1) e (2).<sup>11</sup>

### 4. Potenciais problemas nos modelos tradicionais

Várias são as possíveis candidatas a variáveis excluídas e não linearidades possíveis nos modelos estimados na seção 2. A própria especificação da equação estimada cujos resultados aparecem na tabela 2 já é uma tentativa de superar essa dificuldade, incluindo

---

<sup>11</sup> O coeficiente  $\delta_1$  foi mantido constante nas duas equações, a que representa o processo de geração dos dados e na estimada, mas isso só vai ser verdadeiro em condições especiais, sendo ela que  $E(X_1 X_2) = 0$ . Esse problema foi deixado de lado por não fazer parte da preocupação principal.

variáveis adicionais. Entretanto, nossa hipótese é que algumas variáveis não incluídas, mesmo após essa expansão, podem ser importantes na teoria que serve de base para a Equação de Mincer e podem afetar os resultados referentes à *dummy* regional. Além disso, algumas das variáveis podem também estar sendo incluídas de forma inadequada nas regressões da seção 2 e em vários estudos que recorrem a tais regressões para identificar as desigualdades regionais brasileiras.

Para a melhor compreensão dessas hipóteses, apresentar-se-á uma dedução simples da Equação de Mincer para entender as suposições que ela faz e com isso poder avaliar como as potenciais violações no mundo real podem afetar sua estrutura e os resultados regionais relevantes para esse trabalho. A dedução a ser apresentada será aquela que Mincer (1974) utilizou, havendo outra que apareceu em Mincer (1958), mas que é mais restrita e por tal menos adequada a nossos propósitos.<sup>12</sup>

Sendo  $P_t$  o rendimento potencial de um indivíduo no tempo  $t$ , ele pode dedicar uma parte dele  $k_t$  ( $0 \leq k_t \leq 1$ ) a educação formal ou treinamento. Se  $\rho_t$  for o retorno desse investimento, então é possível dizer que:

$$P_t = P_{t-1}(1 + k_{t-1}\rho_{t-1}) \equiv \prod_{j=0}^{t-1} (1 + \rho_j k_j) P_0 \quad (7)$$

Supor-se-á que no período escolar o indivíduo dedica 100% do seu tempo útil aos estudos e  $k_t=1$  quando ele está na escola. Se chamarmos o retorno da educação formal de  $\rho_S$  e supusermos que ela ocorre no início da vida do indivíduo, a equação (7) pode ser reescrita como:

$$\ln P_t = \ln P_0 + S \ln(1 + \rho_S) + \sum_{j=S}^{t-1} \ln(1 + \rho_0 k_j) \quad (7')$$

Onde se supôs também que os retornos da educação formal  $\rho_S$  e dos treinamentos posteriores  $\rho_0$ , são constantes para todos os anos. Essa equação foi reduzida por Mincer por aproximação, válida apenas para valores baixos de  $\rho_S$  e  $\rho_0$ , gerando:

$$\ln P_t \approx \ln P_0 + \rho_S S + \rho_0 \sum_{j=S}^{t-1} k_j \quad (7'')$$

Mincer (1974) utilizou o modelo de Ben-Porath (1967) para definir a taxa  $k_j$  de investimento em treinamento e aprendizado no trabalho no período pós-escola. Assim ele definiu:

$$k_{S+x} = \kappa \left(1 - \frac{x}{T}\right) \quad (8)$$

<sup>12</sup> A dedução a ser apresentada aqui segue Heckman, Lochman e Todd (2007).

Onde  $x=t-S \geq 0$  é a experiência quando o indivíduo tem a idade  $t$  e  $T$  é o tempo de vida do indivíduo, que se assume ser independente da escolaridade. Definindo-se o rendimento do trabalho observado como o potencial expresso na equação (7'') menos o investimento em treinamento, definido na equação (8), e resolvendo-se o somatório tem-se:

$$\ln w_t \approx (\ln P_0 - \kappa) + \rho_S S + \left( \rho_0 \kappa + \frac{\rho_0 \kappa}{2T} + \frac{\kappa}{T} \right) x - \frac{\rho_0}{2T} x^2 \quad (9)$$

Essa equação põe o logaritmo natural do rendimento do trabalho no tempo  $t$ ,  $\ln w_t$ , como sendo igual a uma constante ( $\ln P_0 - \kappa$ ) mais o retorno marginal de um ano de estudo  $\rho_S$  vezes os anos de estudo do indivíduo,  $S$ , mais uma constante vezes os anos de experiência  $x$ , medida como a idade menos os anos de escolaridade, menos uma constante vezes o quadrado dos anos de experiência.

Essa equação é válida para um indivíduo ao longo de sua vida. Os parâmetros para cada um deles podem não ser iguais aos dos demais. Por exemplo, alguns atributos podem fazer com que o rendimento potencial que o indivíduo perceba inicialmente quando tome suas decisões de escolaridade e treinamento sejam diferentes por causa de suas circunstâncias familiares e físicas. Consequentemente, o termo  $\ln P_0$  pode não ser o mesmo.

Mincer estendeu esse modelo para o caso em que os coeficientes variam de indivíduo para indivíduo, mas com uma variância finita. Nesse caso, a equação (9) torna-se:

$$\ln w_i = \bar{\alpha} + \bar{\rho}_S S_i + \bar{\beta}_0 x_i - \bar{\beta}_1 x_i^2 + e_i \quad (10)$$

Onde:

$$e_i = (\alpha_i - \bar{\alpha}) + (\rho_{Si} - \bar{\rho}_S) S_i + (\beta_{0i} - \bar{\beta}_0) x_i - (\beta_{1i} - \bar{\beta}_1) x_i^2 \quad (11)$$

Várias suposições foram feitas para chegar a esse resultado e tornar essa equação estimável. Uma importante, que gerou toda uma literatura, foi a suposição de que  $\rho_S$  é independente da escolaridade  $S$ . Quando o indivíduo tem mais habilidade, ele tende a ter mais retorno da educação e a permanecer por mais tempo no estudo formal. Assim  $E(\rho_S S) \neq 0$ . Esse problema faz com que haja viés na estimação do retorno da educação.<sup>13</sup> Ele, contudo, não afeta os objetivos desse trabalho e o método a ser apresentado não se abala por essa distorção.

Uma segunda suposição cujo relaxamento pode gerar alterações na forma funcional dessa equação é que o retorno  $\rho_S$  da educação formal é constante para todos os níveis de escolaridade. Essa suposição viola a ideia de retornos marginais decrescentes em Economia e tem se mostrado empiricamente inadequada.<sup>14</sup> Estudos empíricos para os Estados Unidos têm indicado que a relação entre logaritmo natural do rendimento do trabalho e a escolaridade é melhor representada por uma função quadrada na escolaridade. Esse resultado pode ser obtido, por exemplo, na derivação apresentada acima caso se suponha que a taxa de retorno

<sup>13</sup> Esse problema foi destacado por Grilliche (1977) e gerou toda uma literatura que busca bons instrumentos para estimar o retorno da educação por variáveis instrumentais. Ver Card (1999) e Heckman, Lochman e Todd (2007) para discussão desse assunto.

<sup>14</sup> Ver Lamieux (2006) e Belzil (2006).

do investimento em educação cai com a quantidade de anos de estudo. Mais especificamente, caso se introduza a hipótese de que:

$$\rho_j = \rho_s \left( \frac{S-j}{V} \right) \quad (12)$$

Em que V é um número fixo, independente da escolaridade, tal que  $V > S$ . Nesse caso o retorno começa maior, mas decai quando os anos de estudo j aumentam. A equação de Mincer sob essas condições se torna:

$$\ln w_i = \bar{\alpha} + \bar{\rho}_{S1} S_i + \bar{\rho}_{S2} S_i^2 + \bar{\beta}_0 x_i - \bar{\beta}_1 x_i^2 + e_i \quad (10')$$

A equação (11) também ganha mais um termo nesse caso, referente ao valor médio de  $\bar{\rho}_{S2}$ .

$$e_i = (\alpha_i - \bar{\alpha}) + (\rho_{S1i} - \bar{\rho}_{S1}) S_i + (\rho_{S2i} - \bar{\rho}_{S2}) S_i^2 + (\beta_{0i} - \bar{\beta}_0) x_i - (\beta_{1i} - \bar{\beta}_1) x_i^2 \quad (11')$$

Uma terceira suposição que se fez foi que os preços não se alteram ao longo do tempo. Se essa suposição for relaxada, a equação (7) tornar-se-ia:

$$\frac{P_t}{(1 + \pi_{t-1}) P_{t-1}} = \frac{P_{t-1}}{P_{t-1}} (1 + k_{t-1} \rho_{t-1}) \quad \text{ou} \quad P_t = \prod_{j=0}^{t-1} (1 + \pi_{t-j}) (1 + \rho_j k_j) P_0$$

Onde  $p_t$  é o preço no tempo t e  $\pi_t$  é a inflação no tempo t. Caso faça-se  $p_0=1$ , a repetição de toda a derivação entre as equações (7) e (10') gera:

$$\ln w_i = \bar{\alpha} + \ln p_i + \bar{\rho}_{S1} S_i + \bar{\rho}_{S2} S_i^2 + \bar{\beta}_0 x_i - \bar{\beta}_1 x_i^2 + e_i \quad (10'')$$

Onde  $p_i$  nesse caso é um índice de preços que mede o custo de vida a que o indivíduo i está sujeito. Obviamente, quanto maior ele for, maior será o rendimento nominal do trabalho que ele terá.

A observação da equação (10''), mostra que há pelo menos seis fontes básicas de variações entre indivíduos nas relações que ela captura. Sendo i e j dois indivíduos diferentes:

- i.  $\ln W_{0i} \neq \ln W_{0j}$ : Diferenças no rendimento potencial inicial esperado.
- ii.  $\rho_{S1i} \neq \rho_{S1j}$  e  $\rho_{S2i} \neq \rho_{S2j}$ : Diferenças no retorno esperado da educação.
- iii.  $\ln p_i \neq \ln p_j$ : Diferenças nos custos de vida a que estão sujeitos.
- iv.  $\rho_{0i} \neq \rho_{0j}$ : Retorno esperado do treinamento e tempo investido em qualificação.
- v.  $k_i \neq k_j$ : Preferências por treinamento.
- vi.  $T_i \neq T_j$ : Expectativa de vida.

Essas fontes podem ser reunidas em três grupos, que serão discutidos a seguir.

#### 4.1. Atributos do local de residência

Algumas das fontes de diferenciação da equação (10'') entre indivíduos estão associadas ao seu local de residência. O custo de vida ( $\ln p$ ) é o mais óbvio deles. Cada município tem seu próprio custo de vida, que depende entre outras coisas do preço da terra. A expectativa de vida dos indivíduos ( $T$ ) também depende do local de residência, já que cada município e mesmo bairro tem condições específicas de saúde, incluindo-se aí a disponibilidade de saneamento e seu impacto no controle de doenças, por exemplo. As disponibilidades de estrutura e aptidão do ambiente para práticas esportivas, que estão entre os determinantes relevantes da expectativa de vida, também são definidos a nível de município e às vezes até de bairros.

Para capturar esse impacto, sem eliminar a possibilidade de que haja desigualdades de renda entre indivíduos geradas pelas suas regiões de residência, incluíram-se duas variáveis na versão estimada da equação (10''). A primeira foi a população do município de residência. Ela serve como indicador da pressão de demanda sobre o uso do espaço urbano, o que afeta o preço da terra e por consequência o componente dos custos de aluguéis no custo de vida.

A segunda foi os anos de estudos médios da população no município. Ela também ajuda a capturar as diferenças de custo de vida com os quais os indivíduos se defrontam. Cidades com população mais educada tendem a ter maior demanda por espaço, já que eles tendem a ter maior renda. Não se colocou a renda média do município ao invés dos anos médios de estudo porque nesse caso estaríamos explicando as diferenças de renda entre indivíduos a partir da renda per capita da região onde eles vivem. A ideia é testar a hipótese de que as diferenças de rendas individuais e a agregada para as regiões advêm apenas de atributos dos indivíduos, incluindo escolaridade entre eles. Assim, faz sentido a inclusão dos anos médios de estudo na determinação da renda individual, mas não a renda média no município, pois essa última variável transformaria o argumento em uma tautologia, enquanto a primeira não.

Como não há forma funcional definida entre essas duas variáveis e os preços e demais variáveis específicas dos municípios nesse modelo, optou-se pela inclusão dessas variáveis em nível e no seu logaritmo natural. Assim, assegurou-se mais flexibilidade da forma funcional, pois dessa forma tanto a primeira como a segunda derivada da relação podem ser tanto positivas como negativas.

#### 4.2. Atributos relacionados à capacidade individual circunstancial

Os atributos relacionados à capacidade individual circunstancial são aqueles que dependem do indivíduo, mas suas circunstâncias sociais afetam em muito seu desempenho nessas fontes básicas de variações entre indivíduos dos parâmetros da equação (10''). Nas fontes acima mencionadas, o rendimento potencial inicial é o principal deles. Se o indivíduo nasce em uma família cuja inserção social tende a reduzir seu potencial de rendimento futuro,  $\ln P_0$  vai ser menor para esse indivíduo. Além disso, se os indivíduos nascem em ambiente que tem uma certa expectativa sobre ele, normalmente ele tende a formar suas expectativas tomando em consideração aquelas expectativas. Consequentemente, uma mulher em uma sociedade em que se espera um determinado comportamento dela tende a ter suas expectativas de rendimento potencial próprio afetado por essas expectativas. Da mesma forma, indivíduos que nascem em famílias em que se esperam que eles contribuam para a renda familiar ainda

quando adolescentes também tendem a criar expectativas mais modestas de seu rendimento na idade adulta pois sabem que terão uma formação escolar de pior qualidade.

Sendo esses atributos parcialmente determinados pelas condições sociais, eles também podem ter média diferente para as diversas regiões. Assim, uma *dummy* regional como a que aparece nas estimações apresentadas na tabela (1) e (2) podem estar apenas representando as diferenças nas médias regionais de variáveis como  $\ln P_0$ , dadas as determinações a partir de comportamentos sociais como os discutidos aqui.

Entretanto, vale salientar que não é apenas o  $\ln W_0$  que é afetado pelas circunstâncias individuais. As expectativas de retorno da educação e de treinamentos, assim como as preferências por treinamento e mesmo expectativas de vida podem ser afetadas por circunstâncias individuais. Além disso, todas essas circunstâncias individuais podem ter médias regionais diferenciadas por decorrência das condições específicas de ocupação das diversas regiões brasileiras, como retratado em Barros (2011b).

Não havendo informações sobre condições circunstanciais individuais, algumas proxies foram utilizadas aqui, que foram: (i) raça declarada pelo indivíduo; (ii) sexo; (iii) setor de atividade em que se insere, formal ou informal; e (iv) residência em região metropolitana. Essas variações tendem a capturar várias diferenças em circunstâncias individuais e por isso são incluídas com tanta frequência como *dummies* nas estimações da equação de Mincer.

O procedimento adotado aqui, contudo, foi um pouco diferente do que a simples inclusão de *dummies* para essas diferenças óbvias, como é a prática comum. Esse procedimento supõe uma forma funcional muito específica na inclusão dessas variáveis e não considera os seus efeitos nos demais parâmetros do modelo que não seja o  $\ln W_0$ . Por isso, ele foi considerado insuficiente nesse trabalho. Ao invés dessa inclusão de *dummies*, recorreu-se à divisão da amostra para todos esses atributos. Como consequência, permitiu-se que todos os parâmetros da equação (10'') fossem afetados por essas diferenças.

Vale lembrar que a raça foi incluída não porque de fato haja necessariamente diferenças de expectativas associadas a ela. Como bem observou Arrow (1998), ela tende a sinalizar para o mercado de trabalho e realmente refletir circunstâncias de formação doméstica e mesmo de qualidade de educação diferentes. Por isso, essas *proxies* de cor da pele podem capturar diferenças em  $\ln P_0$  e demais parâmetros. A residência em regiões metropolitanas busca capturar as diferenças que existem de expectativas de qualidade de vida e desempenho profissional entre indivíduos de cidades menores e cidades maiores. Muitos indivíduos que migram para as cidades grandes o fazem em busca de mais oportunidades de ascensão social. Isso dá um diferencial no comportamento deles no mercado de trabalho e por tal afeta todos os parâmetros da equação (10'').

### **4.3. Atributos relacionados à capacidade individual**

Os atributos relacionados à capacidade individual são aqueles que dependem apenas do indivíduo e, embora também estejam sujeitos a influência do meio, os indivíduos têm maior capacidade de definir suas situações. Por exemplo, o retorno esperado da educação depende muito do esforço e empenho individual em tornar sua educação de melhor qualidade. Da mesma forma, o retorno esperado dos treinamentos, o percentual de tempo em que o indivíduo se dedica a eles e suas preferências por lazer e mais renda presente ou tal empenho

em cada fase de sua vida profissional dependem muito de decisões individuais. Assim, os indivíduos afetam esses coeficientes e os parâmetros associados a eles na equação (10”) por decisões próprias, que podem ser socialmente determinadas.

As circunstâncias sociais que determinam esses parâmetros podem variar de região para região a partir de diferenças em valores morais e visão de mundo; diferenciadas pelas formações históricas de cada região. Com isso, os parâmetros para as diversas variáveis, incluindo os retornos da educação e da experiência, serão permitidos variar entre regiões nas estimações a serem apresentadas a seguir.

## 5. Método de estimação

A partir dos comentários prévios optou-se por estimar uma equação como a (1) apresentada acima, mas com a seguinte especificação:

$$\ln w_i = \bar{\alpha} + \alpha_1 P + \alpha_2 \ln P + \alpha_3 E + \alpha_4 \ln E + \alpha_5 N + \alpha_6 N^2 + \bar{\rho}_{S1} S_i + \bar{\rho}_{S2} S_i^2 + \bar{\beta}_0 x_i - \bar{\beta}_1 x_i^2 + e_i \quad (13)$$

Onde:

w = rendimento do trabalho.

P = população do município de residência.

E = anos médios de estudo da população do município de residência.

N = Nota do ENEM em 2005 dos estudantes que o fizeram nesse ano.

S = anos de estudo efetivo do indivíduo.

x = anos de experiência, definido como idade menos anos de estudo efetivo menos 6.

O ln antes da variável representa que ela está no seu logaritmo natural. Além da especificação da equação estimada, vale destacar:

- i. Permitir-se-á no modelo estimado a existência de diferenças nos retornos da educação e da experiência nas duas regiões para as quais se busca testar a existência de diferenças de rendas individuais devido ao fator regional.
- ii. Todos os indivíduos poderão ter coeficientes distintos para todas as variáveis incluídas, sendo as estimações apenas para médias para conjuntos específicos deles. Para tornar o método robusto, permite-se que essas médias sejam regionais e não necessariamente sejam iguais nas duas regiões. Isso certamente diminui o impacto potencial do tipo de erro nessas estimações que surgem por causa das não linearidades nas funções que são aproximadas por relações lineares, que foram analisadas na seção anterior.
- iii. Além de possibilitar a diferença de médias de coeficientes para as duas regiões consideradas, permitiu-se também a diferenciação desses coeficientes por sexo na medida em que os resultados que aparecem a seguir são apenas para os homens. Na nossa sociedade as expectativas para homens e mulheres ainda são muito diferentes.

Por consequência, a proporção diferente de mão de obra ativa dos dois gêneros na população pode viesar os resultados se essas diferenças não forem consideradas. Além disso, as perspectivas podem ser diferentes nas duas regiões por diferenças em fatores culturais.

- iv. Também permitiu-se que houvesse diferenciação de expectativas, e por consequência de coeficientes, por raça. Assim restringiu-se a amostra para pardos e brancos apenas e separou-se esses dois grupos em duas amostras.
- v. Permitiu-se que houvesse diferença de expectativas e de coeficientes nas regiões metropolitanas e nos demais municípios. Na verdade essa diferença também decorre da baixa credibilidade das proxies utilizadas para custo de vida, que foi basicamente população e talvez parcialmente anos médios de estudo no município. Sendo essas proxies pobres, provavelmente estão sujeitas a relações não lineares com custo de vida e por isso podem levar a grandes distorções nos extremos de qualquer relação paramétrica, mesmo que não linear. Como as regiões metropolitanas de São Paulo e Rio de Janeiro são muito grandes, os seus habitantes foram excluídos da amostra por provavelmente introduzirem distorções por representação inapropriada do custo de vida nelas. Vale salientar que cidades como Recife, Salvador e Fortaleza no Nordeste continuaram encontrando contra-partes muito próximas no Sul/Sudeste, que são Belo Horizonte, Porto Alegre e Curitiba. Na verdade, a tabela 4 mostra que ordenando essas regiões metropolitanas em cada uma das regiões e criando as contrapartes individuais a partir da posição de ordenação, as três do Sul/Sudeste ainda são maiores do que as três do Nordeste.
- vi. Como a renda na área rural é de difícil mensuração, todos os habitantes da área rural também foram excluídos.
- vii. Também se restringiu a amostra a indivíduos empregados com e sem carteira que não eram funcionários públicos. Assim, excluíram-se da amostra os empregados por conta própria, empregadores e funcionários públicos. As diferenças de expectativas entre esses diversos indivíduos, junto com as diferenças em proporção deles nas populações das diversas regiões e cidades de portes distintos também dificulta a obtenção de médias para regiões, podendo introduzir distorções nos resultados.
- viii. Criou-se outra partição na amostra entre sem carteiras e com carteira. Também nesse caso a ideia é homogeneizar os grupos comparados.

Tabela 4  
População das Regiões Metropolitanas (RM's) cujos habitantes foram incluídos na amostra a ser utilizada nas estimações

RM's	2000	2010	RM's	2000	2010
Recife	3.337.565	3.690.547	Belo Horizonte	4.349.425	4.874.042
Salvador	3.021.572	3.458.571	Porto Alegre	3.718.778	3.958.985
Fortaleza	2.984.689	3.530.942	Curitiba	2.726.556	3.129.269
<b>Total</b>	<b>9.343.826</b>	<b>10.680.060</b>	<b>Total</b>	<b>10.794.759</b>	<b>11.962.296</b>

Fonte: IBGE. Extraído de IPEADATA.

Todas essas partições e exclusões mencionadas acima têm um intuito único, que é tornar mais homogênea a base de comparação e com isso evitar que diferenças regionais sejam geradas de forma espúria por causa de proporções diferentes de tipos de trabalhadores em cada região ou por diferenças culturais que geram diferenças em expectativas apenas. O que se busca verificar é se indivíduos com as mesmas expectativas e mesmos atributos individuais tendem a ter os mesmos níveis de renda nas duas regiões. Assim, distorções não



podem ser criadas em decorrência de proporções diferentes de tipos distintos de trabalhadores nas duas regiões.

A partir dessas partições criaram-se as amostras que aparecem especificadas na tabela 5. Incluiu-se aí dados para ocupados e total apenas no Nordeste porque o método a seguir demanda estimações apenas para esses grupos de trabalhadores. O total inclui os indivíduos que são ativos, mas não estavam trabalhando no mês de referência do Censo de 2000. Esses indivíduos foram incluídos em um primeiro estágio das estimações para gerar a inversa de Mills e num segundo estágio para corrigir os resultados para viés de seleção.

Em cada uma dessas categorias existem diferenças da renda do trabalho média entre as duas regiões, como pode se ver na tabela 6. Pelos desenvolvimentos passados isso pode ocorrer por causa de diferenças em atributos individuais ou dos municípios onde os indivíduos residem ou por causa de diferenças regionais de rendimento. A relevância dessa terceira possibilidade é o que se quer verificar aqui.

Tabela 5  
Amostras por região das categorias estudadas  
(somente homens foram incluídos)

Categoria	Amostra NE	Ocupados no NE
Pardos não metropolitanos com carteira	152.578	138.993
Branco não metropolitanos com carteira	90.932	84.152
Branco metropolitanos com carteira	56.218	54.905
Pardos não metropolitanos sem carteira	82.571	80.562
Branco metropolitanos sem carteira	13.038	12.784
Pardos metropolitanos com carteira	84.594	82.611
Branco não metropolitanos sem carteira	37.120	36.177
Pardos metropolitanos sem carteira	25.169	24.744
Total	542.220	514.928

Fonte: Construída a partir dos microdados do Censo de 2000

Tabela 6  
Rendimentos médios do trabalho nas duas regiões analisadas para os diversos grupos de trabalhadores apresentados na tabela 4 (R\$/mensal)

Categoria	Nordeste	Sudeste	NE/SE
Pardos não metropolitanos com carteira	420,16	492,85	0,85
Branco não metropolitanos com carteira	697,38	921,74	0,76
Branco metropolitanos com carteira (Sem RJ+SP)	1.228,43	1.104,80	1,11
Pardos não metropolitanos sem carteira	209,71	267,98	0,78
Branco metropolitanos sem carteira (Sem RJ+SP)	548,36	587,55	0,93
Pardos metropolitanos com carteira (Sem RJ+SP)	580,21	553,92	1,05
Branco não metropolitanos sem carteira	286,10	402,47	0,71
Pardos metropolitanos sem carteira (Sem RJ+SP)	315,88	266,12	1,19

Fonte: Estimada dos microdados do Censo de 2000.

A equação (13) para cada uma dessas categorias foi estimada no segundo estágio por três métodos. O primeiro foi o de Mínimos Quadrados Ordinários (MQO) com correção para

heteroscedasticidade pelo método de White (1980). O segundo foi o Mínimos Quadrados Generalizados (MQG) com a população dos municípios sendo utilizada como um peso para as variâncias. O terceiro método de estimação foi por Mínimos Desvios Absolutos (MDA).

A tabela 7 traz a estimação por MQG dos coeficientes para um dos segmentos apresentados nas tabelas 5 e 6. Nas estimações utilizou-se uma estrutura de *seemingly unrelated regression* (SUR). Particularmente, o modelo refere-se ao segmento de pardos não metropolitanos que têm certa de trabalho assinada. Essa é a categoria com maior número de trabalhadores no Nordeste, diferentemente do Sul e Sudeste, regiões nas quais a categoria não metropolitanos brancos com carteira é maior. Os números apresentados mostram que os impactos nas rendas individuais do trabalho das diversas variáveis diferem bastante nas duas regiões.

Tabela 7  
Estimações por MQG para a categoria de trabalhadores  
pardos não metropolitanos com carteira

Variável	Nordeste			Sul/Sudeste		
	Coef.	Estat. t	Nível de significância	Coef.	Estat. t	Nível de significância
Constante	-2,16186	-16,963	0,00000			
Anos de estudo	0,12076	36,660	0,00000	0,06197	21,080	0,00000
Anos de estudo ao quadrado	0,00057	2,294	0,02177	0,00248	11,315	0,00000
População do município	0,00000	-4,552	0,00001	0,00000	0,852	0,39430
log natural da população do município	0,05167	7,234	0,00000	0,01436	3,053	0,00227
Anos de estudo médio do município	0,28831	6,370	0,00000	0,06900	1,884	0,05961
Log natural dos anos de estudo médio do município	-0,84284	-6,266	0,00000	0,23646	1,490	0,13625
Experiência	0,05640	50,295	0,00000	0,05379	66,703	0,00000
Experiência ao quadrado	-0,00071	-37,382	0,00000	-0,0008	-48,87	0,00000
Nota do ENEM 2005	0,09636	14,487	0,00000	0,02023	2,915	0,00355
Nota do ENEM ao quadrado	-0,00123	-13,266	0,00000	-0,0002	-2,272	0,02308
Inversa de Mills	0,31708	7,525	0,00000	1,43026	12,408	0,00000
Dummy regional (Sul/Sudeste=1)				1,54485	8,075	0,00000

Fonte: Estimações próprias.

Como as estimações foram feitas em SUR, foi possível se efetuar com facilidade os testes para a hipótese  $H_0$  de que os coeficientes para a mesma variável são iguais. A tabela 8

traz esses testes para os coeficientes que se referem à mesma variável. Como pode se ver, a hipótese de que os coeficientes são iguais nas duas regiões é rejeitada para todas as variáveis com o MQG e quase todos para os dois métodos alternativos. Esse comportamento foi padrão para todas as categorias de trabalhadores. Isso significa que as expectativas e demais fontes de variações de coeficiente entre os indivíduos, apresentadas acima, realmente têm médias diferentes nas duas regiões. Consequentemente, comparar coeficientes para *dummies* regionais não é uma forma adequada de testar a existência de desigualdades regionais de rendas entre indivíduos com atributos semelhantes, como havia sido argumentado anteriormente.

Enfrentando problema semelhante Oaxaca (1973) propôs uma estratégia alternativa para comparar a renda de dois grupos quaisquer, no caso dele, pessoas de sexo diferentes. No método dele, faz-se a comparação de indivíduos médios de uma determinada categoria com retornos diferentes. Essa estratégia, torna-se particularmente problemática quando as equações de determinação das rendas individuais têm relações não lineares como a apresentada na equação (13) e a distribuição dos indivíduos ao longo de qualquer uma das variáveis apresenta *skewness* não desprezível.

Tabela 8

Testes Qui-Quadrado e F da hipótese de que os coeficientes para cada uma das regiões são iguais nas duas regiões com o modelo para pardos não metropolitanos com carteira

Variável	MQG		MQO		MDA	
	$\chi^2$	Signif.	$\chi^2$	Signif.	F	Signif.
Constante	177,28	0,00000	298,70	0,00000	196,783	0,00000
Anos de estudo	33,58	0,00000	12,84	0,00034	0,190	0,66292
Anos de estudo ao quadrado	17,85	0,00002	57,14	0,00000	26,034	0,00000
População do município	19,02	0,00001	95,08	0,00000	61,171	0,00000
log natural da população do município	14,19	0,00017	1,18	0,27788	4,683	0,03046
Anos de estudo médio do município	26,91	0,00000	21,03	0,00000	43,205	0,00000
Log natural dos anos de estudo médio do município	3,59	0,05824	18,59	0,00002	19,293	0,00001
Experiência	7,92	0,00490	13,45	0,00025	134,645	0,00000
Experiência ao quadrado	62,75	0,00000	85,79	0,00000	52,536	0,00000
Nota do ENEMem 2005	66,12	0,00000	87,01	0,00000	58,158	0,00000
Nota do ENEM ao quadrado	82,27	0,00000	267,57	0,00000	454,309	0,00000

Fonte: Elaboração própria.

Diante de tal problema, utilizou-se um método diferente nesse trabalho, mas motivado pela lógica de Oaxaca. Calculou-se os valores médios de cada uma das variáveis nas duas regiões e obteve-se as proporções para cada uma delas entre o Nordeste e o Sul/Sudeste. Essas proporções foram utilizadas para corrigir os valores das variáveis de cada indivíduo no Nordeste de forma que a média da variável para todos os indivíduos nessa região, após tal correção, tornar-se-ia, por construção, igual à média no Sudeste. Após esse estágio, utilizou-se um conjunto dessas variáveis corrigidas e substituiu-se na equação (13) estimada somente com dados do Nordeste. Apenas a constante e os erros em cada indivíduo do Nordeste permaneceram os mesmos estimados originalmente. Após esse estágio, obteve-se a renda

média do trabalho para todos os indivíduos do Nordeste sob essa hipótese alternativa; de que os indivíduos têm os atributos individuais e de seus municípios cujas médias para todos eles eram iguais àquelas encontradas no Sudeste. Por fim, comparou-se essa média da renda do trabalho gerada com essa simulação àquela encontrada para os indivíduos do Sul-Sudeste.

Estão apresentados na tabela 9 os resultados desse exercício para as correções de apenas: (i) anos de estudo, (ii) anos de estudo ao quadrado; (iii) nota média do ENEM no município de residência; (iv) nota média do ENEM no município de residência ao quadrado. No conjunto essas variáveis representam o capital humano individual, considerando a quantidade e a qualidade dos anos de estudo que os indivíduos tiveram. Por isso esse conjunto foi chamado de “apenas educação individual”. Também apresentaram-se na mesma tabela os resultados para a correção mais ampla em que se incluem, além das variáveis apresentadas, as demais referentes a educação, que são anos médio de estudo no município e seu logaritmo natural. Esse conjunto amplo foi chamado de educação total. Em “educação total e população”, por sua vez, acrescentaram-se também correções para as duas variáveis relacionadas com a população do município. Ou seja, nesse último caso apenas as constantes e as variáveis relacionadas às experiências dos indivíduos não tiveram suas médias corrigidas.

Os resultados indicam que para algumas categorias de trabalhadores a renda no Nordeste é maior do que o Sudeste e vice-versa para outras. Ou seja, a ideia de que existe uma disparidade de renda gerada apenas por consequência da localização do trabalhador some com esse método, podendo inclusive ser invertida. Esses resultados indicam que não há evidência conclusiva sobre a existência de disparidades de rendas individuais que sejam explicadas apenas por região de residência, como conclui outros estudos.

Tabela 9

Proporção das rendas médias individuais nos Nordeste e Sul/Sudeste após correções dos atributos individuais e alguns municipais para residentes na primeira região

Categoria	Correções realizadas			
	Original	Apenas educação individual	Educação total	Educação total e população
Homens pardos não metropolitanos com carteira	0,85	0,99	1,15	1,15
Homens brancos não metropolitanos com carteira	0,76	0,96	1,03	1,03
Homens brancos metropolitanos com carteira (Sem RJ+SP)	1,11	0,96	0,95	0,90
Homens pardos não metropolitanos sem carteira	0,78	0,85	0,92	0,90
Homens brancos metropolitanos sem carteira (Sem RJ+SP)	0,93	0,87	0,93	0,92
Homens pardos metropolitanos com carteira (Sem RJ+SP)	1,05	0,93	0,87	0,83
Homens brancos não metropolitanos sem carteira	0,71	0,81	0,81	0,80
Homens pardos metropolitanos sem carteira (Sem RJ+SP)	1,19	1,15	1,17	1,16

Fonte: Estimções próprias a partir dos dados do Censo 2000 e método descrito nessa seção.

Para tentar chegar a uma conclusão mais precisa, optou-se por criar uma medida única que agregue todas as categorias. Essa medida nada mais é do que uma média das proporções para cada uma das categorias incluídas nas tabelas 5 e 6, mas ponderada pela participação da

população ocupada naquela categoria no total do Nordeste. Os resultados mostram que uma proporção para todos os grupos seria de exatamente 1,000. Ou seja, a desigualdade de rendas individuais restante é inexistente caso os atributos individuais relevantes na região Nordeste tivessem médias semelhantes às médias do Sul/Sudeste. As diferenças em cada grupo isoladamente são mais provavelmente consequências de pequenos desvios na representatividade dos dados do que de desigualdades de rendimento do trabalho decorrentes de local de residência.

## 6. Conclusões

As discussões teóricas nesse trabalho mostraram que os métodos tradicionalmente utilizados para concluir que há diferença de rendas individuais do trabalho entre indivíduos morando no Nordeste e no Sul/Sudeste são inadequados. Nesses métodos, as desigualdades encontradas como decorrentes das regiões de residências estão mantendo resquícios de desigualdades decorrentes apenas da composição desproporcional das populações das duas regiões, dado um conjunto de atributos relevantes que geram desigualdades de renda entre indivíduos. Quando esses resquícios são reduzidos ou eliminados, as desigualdades de rendas individuais desaparecem em alguns grupos de indivíduos e caem substancialmente em outros.

O esforço feito aqui para minimizar o impacto de não linearidades nas relações e evitar o efeito de variáveis omitidas e por consequência de reduzir a probabilidade de criar ponderações espúrias que distorçam os resultados não foi totalmente exaustivo quanto às possibilidades existentes para tal. Entretanto, ele já foi suficiente para gerar igualdade de rendimento médio do trabalho entre regiões, quando se considera indivíduos em condições básicas de homogeneidade. Assim, a principal conclusão desse trabalho é que qualquer conclusão de que existe desigualdade de rendas individuais dos trabalhadores decorrentes da região de residência é precipitada. Não há evidência que elas existam que não seja decorrente de erros metodológicos. Ou seja, não há evidência de que o mercado de trabalho brasileiro não seja regionalmente bem arbitrado.

Esse trabalho corrobora com hipóteses, como a apresentada por Barros (2011b), sobre as desigualdades regionais no Brasil. Segundo essa visão elas emergem apenas das desigualdades de capital humano médio entre regiões, não havendo qualquer parte relevante que seja gerada por diferenças em remuneração do mesmo tipo de trabalhador em regiões distintas. Ou seja, não há parte dessas desigualdades regionais que seja originada na falta de arbitragem de remuneração de capital humano entre as regiões.

## Referências

- Arrow, K., "What Has Economics to Say about Racial Discrimination?," *Journal of Economic Perspectives*, 12(2): 91-100, 1998.
- Barros, A., "Decomposição das Desigualdades Regionais Brasileiras em seus Principais Determinantes," Apresentado em ANPEC Regional, Fortaleza, Julho, 2010.
- Barros, A., "The Regional Question in Brazil: Nature, Causes and Policies," em W. Baer e J. Love (eds.), *The Economies of Argentina and Brazil*, London: Edward Elgar, 2011a.
- Barros, A., *Desigualdades Regionais no Brasil: Natureza, Causas, Origens e Solução*, Rio de Janeiro: Elsevier, 2011b.

- Barros, R. P.; M. Carvalho, S. Franco e R. Mendonça, “Determinantes da Queda na Desigualdade de Renda no Brasil.” Texto para Discussão do IPEA, Brasília, 2010.
- Christian Belzil, "Testing the Specification of the Mincer Wage Equation," Working Papers 0608, Groupe d'Analyse et de Théorie Economique (GATE), Centre national de la recherche scientifique (CNRS), Université Lyon 2, Ecole Normale Supérieure, 2006.
- Ben-Porath, “The Production of Human Capital and the Life Cycle of Earnings”, *The Journal of Political Economy*, Vol. 75, No. 4, Part 1, (August):352-365, 1967
- Card, D., "The Causal Effect of Education on Earnings." In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics Volume 3A*. Amsterdam: Elsevier, 1999.
- Cebula, R.J., “Determinants of Geographic Living-Cost Differentials in the United States: An Empirical Note,” *Land Economics*, 56 (November), 476-81, 1980.
- Griliches, Z., "Estimating the returns to schooling: some econometric problems," *Econometrica* 45:1-22, 1977.
- Heckman, J., "Sample Selection Bias as a Specification Error." *Econometrica*, 47(1): 153-161, 1979.
- Heckman, J., L. Lochner e P. Todd, “Earning Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond,” in E. Hanushek e F. Welch (eds.), *Handbook of the Economics of Education*, vol. 1, London: Elsevier, 2007.
- Hoffmann, R., "The Evolution of Income Distribution in Brazil: What Promotes and What Restricts the Decline in Inequality", em W. Baer e J. Love (eds.), *The Economies of Argentina and Brazil*, London: Edward Elgar, 2011.
- Langston D., D.W. Rasmussen e J.C. Simmons, “A Note on Geographic Cost of Living Differentials,” *Land Economics*, 61(3): 314-318, 1985.
- Lee, D. S., and T. Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*, 48(2): 281–355, 2010.
- Lemieux, T., “The “Mincer Equation” Thirty Years After *Schooling, Experience, and Earnings*,” in S. Grossbard (ed.), *Jacob Mincer: A Pioneer of Modern Labor Economics*, New York: Springer, 2006.
- Mincer, J., “Investment in Human Capital and Personal Income Distribution,” *Journal of Political Economy*, 66(4):281-302, 1958.
- Mincer, J., *Schooling, Experience, and Earnings*, New York: NBER Press, 1974.
- Oaxaca, R., “Male-Female Wage Differentials in Urban Labor Markets.” *International Economic Review* 14: 693–709, 1973.

White, H. "A Heteroskedasticity-consistent covariance matrix estimator and direct test for heteroskedasticity". *Econometrica*, 48: 817-838, 1980.