Artigos

Estimando o retorno da educação no Brasil

Edric Martins Ueda§
Rodolfo Hoffmann¤

RESUMO

Este artigo avalia os efeitos da educação sobre os rendimentos individuais no Brasil. Três métodos econométricos são objeto de uma apreciação crítica, bem como as condições necessárias e os limites de cada um para serem aplicados no País. Adicionalmente, analisa-se a influência das condições socioeconômicas da família e das habilidades individuais sobre a determinação das rendas do trabalho. Na parte final do artigo, utilizamos a PNAD-96 (Pesquisa Nacional por Amostras de Domicílios - 1996) com o propósito de estimar as taxas de retorno da escolaridade para as pessoas ocupadas no Brasil. Existem evidências de que, dadas as restrições nos dados disponíveis, esta estimativa está sujeita a apresentar tendenciosidade, o que pode distorcer a compreensão da função que a educação exerce no processo de distribuição de renda.

Palavras-chave: retorno da educação, métodos econométricos, condições socioeconômicas da família, habilidade.

ABSTRACT

This paper evaluates the effects of education on the individual income in Brazil. Three econometric methods are the object of a critical assessment, as well as the necessary conditions and the limitations of their aplication to brazilian data. In addition, the influence of the family background and of the individual abilities on the determination of the labor income are analysed. In the final part of the paper, the PNAD-96 (Nacional Research of Household Samples) data are used to estimate the education return rate for occupied persons in Brazil. There are evidences that, given the restrictions in the available data, this estimative is prone to present bias, that may distort the comprehension of the role that education plays in the process of income distribution.

Key words: return to education, econometric methods, family background, ability.

JEL classification: C31, I21, J31.

---

* Artigo baseado na dissertação de mestrado do primeiro autor, orientada pelo segundo autor. Os autores agradecem os comentários críticos de dois pareceristas desta revista.

§ Mestre em Teoria Econômica pela Unicamp/IE.

¤ Professor da Unicamp/IE.

Recebido em junho de 2001. Aceito em fevereiro de 2002.

1 Introdução

Nos últimos anos vêm sendo feitos importantes progressos na tarefa de estimar os efeitos da educação sobre os rendimentos individuais.[1] Podemos apontar tanto um aperfeiçoamento dos métodos econométricos utilizados como uma melhor compreensão da influência exercida sobre as rendas de trabalho por fatores como as habilidades individuais e as condições socioeconômicas da família. Em especial, este último ponto é importante porque a omissão destes fatores no modelo pode enviesar a estimativa do efeito da educação. O maior problema é que estas variáveis não são facilmente mensuráveis e muitas vezes não são especificadas. Além disso, ainda não existe consenso sobre se os fatores associados às condições socioeconômicas da família exercem, de fato, uma influência significativa sobre os rendimentos individuais.[2]

Geralmente os retornos individuais da educação são estimados por meio de uma equação de rendimento do seguinte tipo (ver, por exemplo, Hoffmann, 2000 e Willis, 1986):

$$Y_i = \alpha + \beta S_i + \sum_j \lambda_j M_{ji} + u_i \qquad (1)$$

onde Y_i é logaritmo do rendimento, S_i representa a variável educação, medida em anos de estudo completos, o coeficiente β é a taxa de retorno da escolaridade, o subscrito i refere-se a um determinado indivíduo e o termo $\sum_j \lambda_j M_{ji}$ engloba as outras variáveis explicativas. Este termo inclui tanto as variáveis que dizem respeito aos atributos produtivos e não-produtivos da pessoa (como a experiência profissional, o sexo e a raça) como aquelas ligadas às características do mercado de trabalho (como a região de moradia e o setor de ocupação).[3]

É importante notar que na equação (1) admitimos uma relação linear entre os anos de estudos completos de uma pessoa e o logaritmo do seu rendimento, mantida constante a experiência no mercado de trabalho. Neste sentido, estamos procurando estimar o efeito de cada ano de escolaridade completo, qualquer que seja o nível educacional, sobre as rendas

1 Neste artigo, quando falarmos em rendimentos estamos nos referindo aos rendimentos obtidos na atividade exercida pela pessoa.

2 Para uma discussão mais detalhada, ver Bowles (1972) e Lam e Schoeni (1993).

3 A origem desta equação pode ser atribuída, por exemplo, à Mincer (1974).

de trabalho. A estimativa b de β obtida é usualmente chamada na literatura de taxa de retorno da educação, apesar de o mais correto ser interpretá-la como a taxa interna privada de retorno, em contraste com a taxa social de retorno. Entretanto, existe uma série de razões que nos estimulam a quebrar esta hipótese, transformando a educação em variáveis binárias destinadas a distinguir os diferentes níveis educacionais (primário, ginásio, colegial e superior).[4] Desta forma, estaríamos considerando que a taxa de retorno da educação não é constante para todo o ciclo escolar, ou seja, que um ano de estudo completo em determinado nível educacional teria um impacto sobre os rendimentos diferente de um ano de estudo em outro nível. Cabe dizer que neste artigo trabalharemos com estas duas formas de especificação. Apesar de a maioria dos trabalhos do gênero utilizarem apenas um modelo básico, análogo à equação (1), o emprego destas duas formas enriquece a análise, permitindo uma pesquisa mais completa.

O método econométrico mais utilizado no Brasil para estimar os efeitos da educação sobre os rendimentos é o bem conhecido método de mínimos quadrados (MMQ).[5] Entretanto, raramente é citado o fato de que as estimativas dos coeficientes da regressão podem sofrer problemas de viés em virtude da (i) omissão de variáveis relevantes no modelo, (ii) de erros de medida na variável educação e (iii) da endogeneidade da escolaridade. Neste sentido, na presença de pelo menos um destes problemas, a taxa de retorno da escolaridade obtida pelo MMQ pode apresentar significativa tendenciosidade. Esta questão não é de pouca importância, já que ela pode distorcer a compreensão do papel que a educação desempenha em temas relacionados à distribuição da renda e à pobreza, tendo em vista que, na grande maioria dos casos, o esforço do pesquisador em avaliar os impactos econômicos desta variável envolve a utilização de microdados e o emprego deste método econométrico.

Em vista destes problemas, uma importante safra de trabalhos tem surgido na literatura internacional mostrando um renovado interesse em oferecer procedimentos econométricos mais adequados para lidar com esta problemática. Entre eles podemos citar os trabalhos de Card (1993, 1999), de Ashenfelter e Rouse (1995) e o de Bound *et al.* (1995). Nestes trabalhos, duas classes de métodos econométricos são os mais utilizados: (i) os que empregam variáveis instrumentais e (ii) aqueles que buscam construir indicadores intrafamiliares (por meio de

4 Pelo fato de existirem diferenças de qualidade no ensino entre os níveis educacionais, forças distintas de oferta e demanda no mercado de trabalho para cada um deles, barreiras à entrada mais fortes no ensino superior do que no fundamental e médio (1°grau e colegial) etc. Entre os trabalhos brasileiros que utilizam esta especificação podemos citar o de Savedoff (1990).

5 Mínimos quadrados ordinários ou ponderados.

amostras de gêmeos univitelinos, irmãos e de pais e filhos) ou que se apóiam em análises de dados de painéis.

O propósito deste artigo concentra-se justamente em analisar estes métodos e investigar em que medida são mais adequados para estimar os efeitos da educação sobre os rendimentos no Brasil, considerando a restrição existente nas bases de dados disponíveis. Adicionalmente, nossa discussão passa pela avaliação da influência das habilidades individuais e das condições socioeconômicas da família sobre as rendas de trabalho.

Este artigo está divido em quatro seções, além desta introdução. Na próxima seção, discutem-se, com mais pormenores, os três problemas que podem surgir ao aplicarmos o MMQ para estimar uma equação semelhante a (1). Na terceira seção são analisados os dois métodos econométricos alternativos utilizados nesta tarefa. Por fim, duas seções se encarregam de analisar a possibilidade de empregá-los no Brasil, utilizando a PNAD-96 (Pesquisa Nacional por Amostras de Domicílios - 1996). Nesta parte, investigamos a robustez dos resultados obtidos pelo MMQ em oposição aos outros métodos econométricos passíveis de serem aplicados e tecemos algumas conclusões sobre as estimativas encontradas.

2 Estimando os retornos da educação pelo MMQ

2.1 Erros de especificação - omissão de variáveis relevantes

Um dos problemas em estimar a equação (1) pelo MMQ é a omissão de variáveis relevantes. Neste caso, é bem conhecido que os estimadores obtidos podem ser viesados e inconsistentes.[6] Isto pode ser facilmente visto num modelo mais simplificado que (1), porém derivado deste. Supondo apenas duas variáveis explanatórias (S_i e A_i) na determinação dos rendimentos individuais, podemos escrever

$$Y_i = \alpha + \beta S_i + \eta A_i + u_i \qquad (2)$$

Se estimarmos a equação (2) pelo MMQ sem considerar a variável A_i, o viés assintótico do estimador b de β é dado por:

6 Mais ainda, os estimadores das variâncias das estimativas dos parâmetros e dos erros também serão viesados.

$$\text{plim}\, b - \beta = \frac{\eta \,\text{cov}(S, A)}{\text{var}(S)} \qquad (3)$$

Se η for positivo, assim como a covariância entre S_i e A_i, b será superestimado. Desta forma, a magnitude e o sinal do viés depende de η e da cov(S, A), ficando claro que se não existir qualquer relação entre a variável omitida e a presente no modelo, o estimador b será não-tendencioso e consistente, mesmo que não se inclua A_i na equação de regressão.

Como já mencionamos, um problema bastante ressaltado nos trabalhos especializados em analisar a relação entre educação e rendimento que se encaixa dentro deste contexto é o fato de não serem incluídas determinadas habilidades individuais ou fatores ligados às condições socioeconômicas da família na equação de rendimento, sob o argumento de que eles afetam **diretamente** as rendas de trabalho (ver, por exemplo, Bowles, 1972). Um trabalho relativamente importante e de interesse para o presente estudo é o de Lam e Schoeni (1993), já que ele analisa a tendenciosidade da taxa de retorno da educação no Brasil provocada pela omissão de variáveis associadas às condições socioeconômicas da família no modelo (com dados extraídos da PNAD-82). A questão é que, como a correlação entre estas variáveis e a educação não é nula, a taxa de retorno da escolaridade fica viesada ao aplicarmos o MMQ. Como argumenta Card (1999), a direção deste viés, em princípio, é desconhecida. Entretanto, a grande maioria dos trabalhos tem encontrado uma tendenciosidade positiva.

Ainda que exista consenso de que as habilidades individuais deveriam ser especificadas no modelo - e a questão passa a ser apenas um problema de encontrar variáveis que mensurem adequadamente estes atributos -, não se pode falar o mesmo para os fatores associados às condições socioeconômicas da família. Neste caso, muitos pesquisadores advogam que no momento da contratação do funcionário, variáveis como o rendimento dos pais ou seus níveis de escolaridade não influenciariam as decisões dos empregadores e, portanto, não haveria motivos para esperar que as condições socioeconômicas da família afetassem diretamente os rendimentos dos filhos. Na verdade, eles só teriam um efeito indireto sobre as rendas de trabalho por meio da educação, já que estes fatores são, na maioria dos casos, os mais importantes na determinação do grau de escolaridade atingido por uma pessoa. Assim sendo, eles não deveriam estar presentes na equação (1). Entretanto, dois motivos nos levam a discordar deste tipo de argumento: primeiro, porque pais melhor qualificados podem ter acesso a oportunidades de emprego com melhor remuneração para seus filhos e, em segundo lugar, pelo fato de que podem ocorrer transferências, dos pais para os filhos, de ativos (empresas, fazenda etc.), heranças e outras coisas semelhantes que são importantes para definir os rendimentos futuros dos herdeiros. Nos trabalhos de Becker (1964) e Jencks (1972) podemos

encontrar uma discussão mais detalhada sobre estes efeitos. Neste sentido, podem existir tanto um efeito direto como um indireto dos fatores associados às condições socioeconômicas da família sobre as rendas de trabalho, o que implica que eles devem ser especificados na equação de rendimento.

Uma razão adicional que nos motiva a incluir no modelo os fatores associados às condições socioeconômicas da família é que eles podem servir como *proxies* de determinadas características de personalidade ou cultura que afetam também diretamente os rendimentos (tais como a persistência, a ambição, a iniciativa e o desembaraço). Cabe dizer que estas características são valorizadas pelo mercado de trabalho porque afetam o desempenho profissional de uma pessoa e são qualidades que não são facilmente "adquiríveis"; dependem do tipo de criação que a pessoa teve, do círculo de amizades que freqüentou, da escola que estudou, enfim, de uma série de fatores responsáveis pelo seu desenvolvimento.[7] O problema todo é que elas não são facilmente mensuráveis, e por este motivo muitas vezes não são passíveis de serem incluídas na equação de rendimento, o que acaba introduzindo uma fonte adicional de viés nas estimativas dos efeitos da educação. Contudo, se admitirmos que o ambiente familiar é um dos fatores que mais influenciam na formação destas características (que, mais tarde, são reforçadas em instituições secundárias como a escola), ele poderá servir como *proxy* destas variáveis à medida que a correlação entre eles for forte. Se assim for, os fatores associados às condições socioeconômicas da família incorporarão parte dos efeitos exercidos pelas características de personalidade sobre os rendimentos, ainda que elas não estejam especificadas no modelo. Dessa maneira, o viés da taxa de retorno da escolaridade pode ser reduzido.

No trabalho de Lam e Schoeni (1993), por exemplo, são incluídas algumas variáveis para avaliar a influência das condições socioeconômicas da família sobre os rendimentos no Brasil, com os autores chegando à conclusão que elas são importantes para explicar a conformação dos rendimentos do trabalho. Entretanto, na visão dos mesmos, o efeito destas variáveis sobre os rendimentos provavelmente se deve mais ao fato de estas serem *proxies* de características pessoais não observáveis do que em virtude de as condições socioeconômicas da família afetarem diretamente os rendimentos por meio de transferência de ativos ou pelo "nepotismo" existente no mercado de trabalho via conexões entre pais e filhos. Cabe dizer apenas que os resultados apresentados corroboram os argumentos aqui apresentados para a inclusão destes fatores no modelo, mas dão margem a interpretações diferentes da do artigo.

7 Não é pequeno o número de pesquisas que apontam o quão decisivos são estes atributos na contratação de pessoas para determinados cargos (ver, por exemplo, Bowles, 1973).

Podemos, no entanto, apontar que caso não sejam especificadas as habilidades individuais e os fatores associados às condições socioeconômicas da família no modelo (por problemas de mensuração ou por questões conceituais), estaremos desconsiderando uma importante dimensão na determinação dos rendimentos individuais. Isto acaba introduzindo um viés nas taxas de retorno da educação, não só por desconsiderar os efeitos diretos destas variáveis, como também por desprezar o fato de que estes fatores podem captar a influência de outros atributos produtivos que não são mensuráveis na prática e que afetam diretamente os rendimentos.

2.2 Erros de medida

Um outro problema que surge quando aplicamos o MMQ para estimar a equação de rendimento é a possibilidade de as variáveis apresentarem erros de medida. Em especial, diversos estudos têm insistido no fato de que a variável educação sofre tal deficiência. Ashenfelter e Rouse (1995), por exemplo, verificaram a existência de erros de medida nos anos de estudo reportados numa amostra de gêmeos, encontrando um viés negativo de aproximadamente 30% na taxa de retorno da escolaridade (β).

Em notação matricial, podemos expor este problema da seguinte forma:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\lambda} + \beta\mathbf{s} + \mathbf{u} \tag{4}$$

$$\mathbf{s}^* = \mathbf{s} + \mathbf{v} \tag{5}$$

onde **u** e **v** são os vetores-coluna dos termos aleatórios das respectivas equações, independentes entre si, **s** representa o vetor da variável não-observável educação, $\mathbf{s}^*$ é o vetor dos valores observados da educação (incluindo o erro de medida), β é um escalar e **X** é a matriz com as demais variáveis explanatórias.

Substituindo (5) em (4), obtém-se uma equação de regressão com a educação observada ($\mathbf{s}^*$) como variável explanatória e erros ($\mathbf{e} = \mathbf{u} - \beta\mathbf{v}$). Como $\mathbf{s}^*$ é correlacionada com **e**, se aplicarmos o MMQ, os estimadores dos parâmetros apresentarão problemas de tendenciosidade. Para erros de medida aleatórios e $\beta > 0$, este viés é inequivocamente negativo.

Uma questão adicional é que o viés provocado pelo erro de medida pode se agravar quando tomado em conjunto com as tentativas de melhorar a especificação do modelo. Como

foi discutido no item anterior, a omissão de variáveis no modelo pode levar à superestimação dos parâmetros. Entretanto, um "excesso de zelo" do pesquisador em se proteger contra este tipo de problema muitas vezes leva-o a incluir variáveis que são irrelevantes. A conseqüência direta disto seria a ampliação do viés negativo causado pelo erro de medida (o que subestimaria ainda mais os parâmetros). Griliches (1977) analisa com elegância esta problemática.

2.3 Endogeneidade da escolaridade

Por fim, o terceiro problema a ser mencionado é a possibilidade de a variável educação ser endógena. Podemos explicá-lo de forma mais apropriada considerando as equações abaixo em notação matricial:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\lambda} + \beta\mathbf{s} + \mathbf{u} \tag{4}$$

$$\mathbf{s} = \mathbf{H}\boldsymbol{\psi} + \mathbf{w} \tag{6}$$

onde **H** é a matriz das variáveis explanatórias da equação da educação e **w** e **u** são os vetores-coluna dos erros aleatórios.

O problema da endogeneidade da educação surge quando **u** e **w** são correlacionados entre si, fazendo com que a variável educação passe a ser também correlacionada com o termo aleatório da equação de rendimento. Isto invalida a aplicação do MMQ, já que é quebrado um dos seus postulados básicos. Caso se proceda à estimação por mínimos quadrados, os estimadores dos parâmetros novamente serão viesados e inconsistentes.

Em princípio, a correlação entre **s** e **u** advém do fato de o termo aleatório incorporar fatores **desconhecidos** que guardam relação com a variável educação e que não são expressos na equação de rendimento. Assim, em termos práticos, a endogeneidade da escolaridade pode ser interpretada como um caso de omissão de variável; entretanto, é uma variável ainda não identificável até o momento.

3 Outros métodos econométricos utilizados para analisar a relação entre educação e rendimentos

A presença destes três problemas na aplicação do MMQ tem estimulado um rico debate sobre quais outros procedimentos poderiam ser empregados na tarefa de estimar os retornos

da educação, com o intuito de se obter estimadores mais consistentes para a taxa de retorno da escolaridade. Passaremos a discutir os dois tipos de métodos citados na parte introdutória deste artigo que vêm sendo mais utilizados.

3.1 Método de estimação por variáveis instrumentais

Quando levamos em consideração os problemas de erro de medida e da endogeneidade da escolaridade, o procedimento indicado para estimar consistentemente os parâmetros da equação de rendimento é a utilização de variáveis instrumentais (MVI) - aplicando, por exemplo, o método de mínimos quadrados em dois estágios ou o método de variáveis instrumentais. Como é bem conhecido, existem duas condições que devem ser satisfeitas para que o instrumento utilizado seja apropriado: que ele seja assintoticamente correlacionado com a variável educação, mas não com o termo aleatório da equação de rendimento. Estas condições estão representadas abaixo pela equação (8) e (9), respectivamente:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{7}$$

$$\text{plim}\left(\frac{1}{n}\mathbf{Z}'\mathbf{s}\right) \neq \mathbf{0} \tag{8}$$

$$\text{plim}\left(\frac{1}{n}\mathbf{Z}'\mathbf{u}\right) = \mathbf{0} \tag{9}$$

onde **Z** representa a matriz das variáveis instrumentais e a equação (7) é o modelo básico da regressão linear múltipla. Note que na expressão (7) a matriz **X** inclui a coluna **s**, diferentemente do que ocorria na expressão (4). Os estimadores obtidos pelo método de variáveis instrumentais são:[8]

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$$

Desta forma, lembrando da equação (4), o viés provocado pela correlação existente entre **s** e **u** é contornado. Entretanto, deve-se deixar assinalado que, na prática, os estimadores do

8 No caso de identificação exata da equação, os mesmos estimadores podem ser obtidos pelo método de mínimos quadrados em dois estágios.

MVI podem não ser os mais apropriados. Para o nosso caso, por exemplo, se a correlação entre a educação e o instrumento for fraca, apenas uma pequena associação entre o termo aleatório da equação de rendimento e a variável instrumental pode produzir estimadores mais inconsistentes que os do MMQ. (Bound *et al.*,1995) Isto pode ser observado tomando como exemplo o seguinte modelo de duas equações:

$$y = \beta s + u \tag{10a}$$

$$s = \pi z + w \tag{10b}$$

onde z é uma variável instrumental já centrada. Considerando os estimadores de β pelo MMQ ($\hat{\beta}_{MQ}$) e pelo método de variáveis instrumentais ($\hat{\beta}_{VI}$), temos o seguinte:

$$\operatorname{plim} \hat{\beta}_{MQ} = \beta + \frac{\operatorname{cov}(s,u)}{\sigma_s^2} = \beta + \frac{\sigma_u}{\sigma_s} \rho_{su} \tag{11}$$

$$\operatorname{plim} \hat{\beta}_{VI} = \beta + \frac{\operatorname{cov}(z,u)}{\operatorname{cov}(z,s)} = \beta + \frac{\sigma_u}{\sigma_s} \cdot \frac{\rho_{zu}}{\rho_{zs}} \tag{12}$$

onde ρ_{ij} é o coeficiente de correlação entre as variáveis i e j. Efetuando algumas transformações nas equações (11) e (12), e dividindo a última pela primeira temos:

$$\frac{\operatorname{plim} \hat{\beta}_{VI} - \beta}{\operatorname{plim} \hat{\beta}_{MQ} - \beta} = \frac{\rho_{zu}}{\rho_{su} \quad \rho_{zs}} \tag{13}$$

Assim, é possível verificar que quanto mais fraca for a correlação entre z e s, ainda que ρ_{zu} seja pequeno, maior tende a ser a inconsistência do estimador de variáveis instrumentais em relação ao obtido pelo MMQ. Vale repetir, contudo, que se ρ_{zu} for nulo, o MVI evita a inconsistência provocada pelos erros de medida na educação ou pela endogeneidade desta variável.

O viés causado pela omissão de variáveis na equação de rendimento, contudo, permanece (no nosso caso refere-se à omissão de variáveis como habilidades específicas ou fatores

associados às condições socioeconômicas da família). A solução dada a esta questão tem sido a utilização cada vez maior de *proxies*. Em relação aos fatores ligados às condições socioeconômicas da família, adotam-se indicadores como o grau de escolaridade dos pais e o trabalho que exerciam na idade escolar de seus filhos. Quanto à variável genérica habilidade, são tomados como *proxies* testes psicotécnicos e de aptidão (QI e outros semelhantes). Entretanto, existem ainda muitas dúvidas sobre até que ponto as *proxies* utilizadas podem mensurar, de forma satisfatória, estas variáveis. No caso da(s) *proxy(ies)* para as habilidades, os problemas podem ser ainda maiores: até hoje não se sabe bem o que estes testes medem e é forçoso admitir que eles não conseguem quantificar, de forma adequada, quais habilidades são natas do indivíduo e quais se devem à educação recebida. Isto é de particular importância para o caso em estudo porque o eventual emprego destas variáveis não ajuda muito a responder em que medida as habilidades individuais condicionam a educação recebida ou, de maneira inversa, em que medida a educação afeta as habilidades individuais.

3.2 Estimadores intrafamiliares e análises intertemporais de um mesmo grupo de pessoas

Outras abordagens para as questões relativas aos erros de medida na educação, da endogeneidade desta variável e da omissão de variáveis relevantes incluem (i) os métodos que buscam construir estimadores intrafamiliares e (ii) aqueles que se munem de séries temporais de dados de um mesmo grupo de indivíduos, combinados ou não com o MVI. A idéia básica desta abordagem é utilizar várias observações de um mesmo indivíduo ao longo do tempo ou informações de duas pessoas que possuem características semelhantes em termos de habilidades, provenientes da mesma família. Desta forma, procura-se isolar os efeitos que as habilidades individuais ou que as condições socioeconômicas da família podem exercer direta ou indiretamente sobre os rendimentos. Duas são as hipóteses por trás destes procedimentos: (i) admite-se que as condições socioeconômicas da família são relativamente iguais para os parentes ou constantes para o mesmo indivíduo ao longo do tempo e (ii) que as habilidades individuais também são bem semelhantes entre os parentes ou constantes para o indivíduo ao longo do tempo (ver Isacsson, 1999).

Um jeito simples de compreender o que foi dito acima é examinar um dos dois modelos apresentados a seguir:

$$Y_{1t} = C_1 + \beta S_{1t} + \sum_j \lambda_j X_{j1} + u_{1t} \qquad (14)$$

$$Y_{1,t+1} = C_1 + \beta S_{1,t+1} + \sum_j \lambda_j X_{j1} + u_{1,t+1} \quad (15)$$

ou

$$Y_{1i} = C_i + \beta S_{1i} + \sum_j \lambda_j X_{ji} + u_{1i} \quad (16)$$

$$Y_{2i} = C_i + \beta S_{2i} + \sum_j \lambda_j X_{ji} + u_{2i} \quad (17)$$

onde (14) e (15) são as equações de rendimento de um mesmo indivíduo em dois momentos distintos do tempo e (16) e (17) são as equações de dois parentes num mesmo instante de tempo. C representa as características da habilidade do indivíduo e as condições socioeconômicas da família que, por suposição, são praticamente as mesmas para o mesmo indivíduo ao longo do tempo ou para os dois parentes, e o termo $\sum_j \lambda_j X_{j1}$ ou $\sum_j \lambda_j X_{ji}$ inclui variáveis explicativas constantes no tempo ou iguais para integrantes da mesma família.[9] Subtraindo (14) de (15) e (16) de (17), obtemos as equações (18) e (19), a partir das quais podem ser calculados os estimadores de β sem problemas de vieses causados pela omissão de variáveis relevantes (como determinadas habilidades individuais e outras ligadas às condições socioeconômicas da família).

$$Y_{1,t+1} - Y_{1,t} = \beta(S_{1,t+1} - S_{1,t}) + u_{1,t+1} - u_{1,t} \quad (18)$$

$$Y_{2i} - Y_{1i} = \beta(S_{2i} - S_{1i}) + u_{2i} - u_{1i} \quad (19)$$

Quanto ao problema da endogeneidade da escolaridade, esta abordagem fornece o seguinte encaminhamento: admite-se que os termos aleatórios das equações de rendimento e educação têm uma estrutura de covariância restrita. Qualquer correlação que possa existir en-

9 Não é necessário que o termo $\sum_j \lambda_j X_{j1}$ ou $\sum_j \lambda_j X_{ji}$ seja constante; só admitimos isto para efeito de simplificação.

tre eles se deve a fatores desconhecidos que estão associados às habilidades não-observáveis e às características socioeconômicas da família não-observáveis. Assim, se nossos dados referem-se a indivíduos da mesma família ou a um mesmo indivíduo, onde as condições (i) e (ii) são obedecidas, a correlação entre os erros é suprimida.

Um dos inconvenientes desta abordagem é que os estimadores obtidos tendem a ser mais sensíveis aos erros de medida na variável educação. Vamos examinar o que ocorre com um exemplo simples. Consideremos inicialmente o caso de dados individuais, com erro na medida da educação:

$$y_i = \beta s_i + u_i \tag{20}$$

$$s_i^* = s_i + v_i \tag{21}$$

Se o estimador b de β for obtido calculando uma regressão linear simples de y_i contra s_i^* (a educação observada com erro de medida), temos:

$$\operatorname{plim} b - \beta = -\beta \frac{\sigma_v^2}{\sigma_{s^*}^2} \tag{22}$$

Vejamos, agora, o caso em que temos dados para pares de indivíduos de uma mesma família:

$$y_{2i} - y_{1i} = \beta(s_{2i} - s_{1i}) + u_{2i} - \overline{u}_{2i} - u_{1i} + \overline{u}_{1i} \tag{19a}$$

$$s_{1i}^* = s_{1i} + v_{1i} \tag{23}$$

$$s_{2i}^* = s_{2i} + v_{2i} \tag{24}$$

Note-se que a equação (19a) já está com todas as variáveis centradas. Se o estimador b de β for obtido calculando uma regressão linear simples de $\Delta y_i = y_{2i} - y_{1i}$ contra $\Delta s_i^* = s_{2i}^* - s_{1i}^*$, temos

$$\text{plim}\, b - \beta = -\beta \frac{\sigma^2_{\Delta v}}{\sigma^2_{\Delta s^*}} \qquad (25)$$

onde $\sigma^2_{\Delta v}$ indica a variância de $\Delta v_i = v_{2i} - v_{1i}$.

Comparando (22) com (25), é bem possível que o viés assintótico da última equação supere o da primeira, já que é razoável esperar que a relação $\sigma^2_{\Delta v}/\sigma^2_{\Delta s^*}$ seja maior do que a relação $\sigma^2_v/\sigma^2_{s^*}$. [10] Conforme apontado na seção anterior, um dos procedimentos aplicáveis neste caso é o MVI, que passa a ser "combinado" com a metodologia desta abordagem.

Em que medida esta abordagem fornece um solução mais satisfatória que a descrita no item anterior vai depender da veracidade das hipóteses por trás destes experimentos ((i) e (ii)). Note-se que é difícil obter dados referentes a várias observações de uma mesma pessoa, com variação na sua escolaridade, como é necessário para estimar β na equação (18).

Em relação às amostras de parentes - pais e filhos, irmãos e gêmeos -, existem problemas também. Os experimentos com pais e filhos são os mais deficientes: as condições socioeconômicas da família podem variar muito entre as duas gerações e as habilidades dos pais e filhos, ainda que sejam semelhantes, não são de forma alguma iguais. Por sua vez, irmãos têm a vantagem de possuir habilidades mais parecidas que entre pais e filhos, pois além da semelhança genética, normalmente desfrutam de ambientes parecidos no seu crescimento. Entretanto, novamente é difícil supor que possuam habilidades exatamente iguais, por razões óbvias. Somente em trabalhos que analisam gêmeos univitelinos é que as hipóteses (i) e (ii) são mais defensíveis. Neste caso, é razoável admitir que, se as diferenças entre as escolaridades dos gêmeos forem acompanhadas de distintos rendimentos, isto se deve predominantemente ao efeito causal e direto da educação (e não de terceiras variáveis), sendo passível de ser mensurada por métodos econométricos.

10 Tendo em vista que $\sigma^2_{\Delta v} = \sigma^2_{v_1} + \sigma^2_{v_2} - 2\,\text{cov}(v_1, v_2)$ e $\sigma^2_{\Delta s^*} = \sigma^2_{s_1^*} + \sigma^2_{s_2^*} - 2\,\text{cov}(s_1^*, s_2^*)$ e que a $\text{cov}(s_1^*, s_2^*)$ tend[illegible]

a $\text{cov}(v_1, v_2)$ pequena.

4 Estimando os retornos da educação no Brasil

4.1 Base de dados

Uma das bases de dados mais completa para analisar os retornos da educação no Brasil é a Pesquisa Nacional por Amostras de Domícilios (PNAD/IBGE), já que é a única de abrangência nacional com informações bastante detalhadas sobre as características dos trabalhos e rendimentos individuais, conjuntamente com dados sobre a educação das pessoas. Em especial, a PNAD-96 inclui, no questionário básico, perguntas referentes aos pais das pessoas de referência nos domicílios e respectivos cônjuges, com o intuito de estudar a mobilidade social. Assim, ela permite investigar a influência que o ambiente socioeconômico familiar exerce sobre a educação e os rendimentos.

Como não é em todos os anos em que se faz a pesquisa que estas informações são coletadas, a PNAD-96 é de particular importância para nosso interesse.[11] Entretanto, cabe dizer que em qualquer das PNADs é possível fazer comparações entre pais e filhos. Contudo, o que normalmente ocorre é que os filhos identificados dessa maneira não estão na *prime age* (25 a 60 anos). Isto acaba desqualificando a utilização desses dados neste artigo, já que, para o propósito desta pesquisa, são imprescindíveis informações sobre os rendimentos dos filhos.

De um total de 331.263 observações disponíveis na PNAD-96, só consideramos as pessoas ocupadas na semana de referência e com declaração bem definida nas variáveis explanatórias que utilizamos, inclusive nas associadas às condições socioeconômicas da família. Adicionalmente, excluímos as pessoas sem rendimento ou que trabalhavam na produção/ construção para o próprio consumo/uso e aquelas cujo pai não estava trabalhando quando tinham quinze anos de idade. É importante notar que as informações referentes às condições socioeconômicas das famílias só são coletadas para as pessoas de referência ou os cônjuges. Desta forma, indivíduos com as demais condições na família (filhos, parentes, pensi) foram desconsiderados. No final, restaram cerca de 55.000 observações. Na t apresentamos algumas características da amostra utilizada.

11 Além da PNAD-96, informações a respeito dos pais t

Tabela 1
Características da Amostra

Escolaridade do Pai	Amostra	Amostra(%)	Característica dos filhos					
			Escolaridade Média	Rendimento médio (R$)	Idade Média	Homens (%)	Branco (%)	Urbano (%)
menos que 1 ano	20,244	36,97	4,55	328,80	41,97	67,34	43,50	72,39
1-4 anos	26,521	48,43	8,12	633,30	38,97	62,18	64,85	85,70
5-8 anos	3,583	6,54	11,10	987,89	36,84	56,13	65,95	96,18
9-11 anos	2,628	4,80	13,01	1.447,83	37,83	54,11	75,23	96,84
12 ou mais anos	1,781	3,25	14,27	2.141,94	37,67	56,99	85,29	96,69
Total	54,757	100,00						

É importante observar a forte correlação existente entre as escolaridades dos pais e dos filhos, além da sensível melhora no nível educacional para a geração dos filhos.[12] Para os quatro primeiros níveis educacionais dos pais, a média dos anos de estudo dos filhos é sempre maior que o limite superior do estrato. Pode-se notar também que quanto mais instruído o pai, maior é a porcentagem de pessoas brancas e que moram em áreas urbanas, assim como maior é o rendimento médio. Estes resultados podem apontar, em princípio, a existência de discriminação e segmentação no mercado de trabalho e ressaltam a importância de incluir essas variáveis na análise da determinação dos rendimentos.

4.2 Modelo

Dois modelos serão utilizados nas estimações:

$$Y_i = \alpha + \beta S_i + \sum_j \lambda_j M_{ji} + u_i \qquad (1)$$

$$Y_i = \alpha + \sum_j \beta_j S_{ji} + \sum_j \lambda_j M_{ji} + u_i \qquad (26)$$

A diferença entre as equações (1) e (26) está na forma de especificar a variável educação. Na equação (1) ela é especificada como uma variável contínua (anos de estudo completos); na equação (26) como quatro variáveis binárias para distinguir cinco níveis de escolaridade

12 Tendência similar é apontada nos trabalhos de Pastore e Silva (2000) e de Lam e Schoeni (1993).

(termo $\sum_j \beta_j S_{ji}$): analfabetos ou com menos de um ano de estudo (tomado como base), primário (1 a 4 anos de estudo), ginasial (5 a 8 anos de estudo), colegial (9 a 11 anos de estudo) e nível superior (12 ou mais anos de estudo). Note-se que as denominações utilizadas não significam que o correspondente nível de escolaridade tenha sido completado, obtendo o respectivo diploma. Assim sendo, as taxas anuais foram calculadas considerando a média dos anos de estudo de cada nível. Por exemplo, se b_1 é a estimativa do coeficiente de regressão da variável binária usada para o primário, a correspondente taxa anual de retorno não é calculada por $T_P = 100[\exp(b_1/4) - 1]$, mas por $T_P = 100[\exp(b_1/2{,}5) - 1]$, pois 2,5 é a média dos anos de estudo deste nível (que compreende o intervalo de 1 a 4 anos). O coeficiente de regressão de uma variável binária para determinado nível de educação não mensura o ganho associado ao nível completo, mas o efeito esperado no rendimento de todas as pessoas que atingiram aquele nível de escolaridade. No caso do nível "primário", por exemplo, tem-se pessoas com 1 a 4 anos de escolaridade. Assim, ao calcular a taxa anual de retorno à educação é necessário considerar a média dos anos de escolaridade correspondente a cada nível, e não o total de anos necessários para completá-lo. As taxas de retorno para os outros níveis de escolaridade obedeceram a seguinte fórmula:

$$T = 100\left[\exp\left(\frac{b_i - b_{i-1})}{(\mu_i - \mu_{i-1})}\right) - 1\right],$$

onde μ_i representa a média dos anos de estudo do nível i e μ_{i-1} é a média dos anos de estudo do nível abaixo. De forma semelhante, b_i é a estimativa do coeficiente de regressão da variável binária do nível i e b_{i-1} é a estimativa do coeficiente de regressão da variável binária do nível abaixo. Cabe dizer que consideramos que são necessários seis anos para concluir o nível superior, o que, obviamente, é uma aproximação, inclusive porque este estrato inclui pessoas com mais de 17 anos de estudo (mestrado, doutorado etc.). Entretanto, pela PNAD-96 não é possível quantificar de forma precisa o limite superior deste nível de ensino. Desta forma, cabe indicar que esta taxa pode estar superestimada.

A variável dependente utilizada é o logaritmo neperiano do rendimento de todos os trabalhos das pessoas ocupadas e as outras variáveis explanatórias (termo $\sum_j \lambda_j M_{ji}$) incluem:

a) variáveis para as características pessoais: quatro variáveis binárias para distinguir a cor; uma variável binária para diferenciar a pessoa de referência; uma variável binária para distinguir o sexo; e duas variáveis para a idade da pessoa (idade e seu quadrado);

b) variáveis representando as características do emprego e do mercado de trabalho: uma variável binária para diferenciar domicílios urbanos; duas variáveis binárias para distinguir três setores de ocupação (agricultura, indústria e serviços); cinco variáveis binárias para regiões; quatro variáveis binárias para distinguir cinco categorias de posição na ocupação (empregado, funcionário público e militar, conta própria, empregador e empregado doméstico, sendo esta última categoria tomada como base); e quatro variáveis binárias para distinguir os intervalos de horas trabalhadas por semana em todos os trabalhos.

Outros três tipos de variáveis explanatórias são incluídas nas equações onde são especificadas as condições socioeconômicas da família:

a) quatro variáveis binárias para a escolaridade do pai e quatro para a da mãe, onde os níveis definidos são os seguintes: nível 5 - superior completo ou mestrado/doutorado completos; nível 4 - de ensino médio incompleto (colegial) a superior incompleto; nível 3 - ginásio completo ou incompleto; nível 2 - primário completo ou incompleto; nível 1 - analfabeto ou com menos de 1 ano de estudo;

b) cinco variáveis binárias para distinguir 6 tipos de grupos de ocupação do pai quando o filho possuía 15 anos: alto (nível 6), médio-superior (nível 5), médio-médio (nível 4), médio-inferior (nível 3), baixo-superior (nível 2) e baixo-inferior (nível 1), sendo o último a base.

Para classificar as ocupações nestes 6 grupos diferentes, foi adotado um critério de *status* socioeconômico, à semelhança do que foi feito por Pastore e Silva no livro *Mobilidade Social no Brasil* (2000). Seguimos as seguintes etapas:

1) Foi criada uma tabela bidimensional com anos de escolaridade completos da pessoa *versus* grupos de idade (*proxy* para a experiência). Para cada célula da tabela foi calculada a renda esperada. A média aritmética entre esta renda esperada e a renda de todos os trabalhos de um indivíduo foi chamada de *status* socioeconômico individual É importante notar que não se trata apenas de *status* econômico, mas de *status* socioeconômico, já que não é levado em conta apenas o rendimento da pessoa, mas também sua escolaridade e idade.

2) Para cada ocupação existente (são 381 na PNAD-96), calculou-se a média aritmética do *status* socioeconômico individual. Este seria o *status* socioeconômico da ocupação (*SSE*).

3) Por fim, calibrou-se a escala do *SSE* de 0 a 100 e foi feita a classificação tendo em vista a distância socioeconômica entre os grupos e a parcela da população contida neles. A tabela abaixo mostra a classificação.

Tabela 2
Grupos de Ocupação

Grupos de Ocupação	Intervalo *SSE*	Média *SSE*	% Pop	Distância Socioec.
1 (baixo-inferior)	$0 \leq SSE < 5$	2,46	38,6	1,0
2 (baixo-superior)	$5 \leq SSE < 10$	6,87	29,2	2,8
3 (médio-baixo)	$10 \leq SSE < 15$	11,71	17,5	4,8
4 (médio-médio)	$15 \leq SSE < 25$	20,00	6,8	8,1
5 (médio-superior)	$25 \leq SSE < 40$	30,82	5,2	12,5
6 (alto)	$40 \leq SSE \leq 100$	49,14	2,7	20,0

É importante notar que a utilização da escolaridade dos pais para retratar as condições socioeconômicas da família é mais comum na literatura. Entretanto, o ideal seria dispor de outras variáveis, como a renda dos pais ou a riqueza familiar, já que a escolaridade capta apenas **uma dimensão** das condições socioeconômicas. Neste sentido, a inclusão dos grupos de ocupação do pai no modelo tem a intenção de complementar a especificação. Lam e Schoeni (1993), por exemplo, analisando dados da PNAD de 1982, consideram a escolaridade dos pais e também as escolaridades dos sogros e do cônjuge. A idéia de incorporar informações a respeito destes últimos familiares provém de modelos econômicos inspirados no "mercado" matrimonial, argumentando-se que aí ocorre uma avaliação das características da pessoa parecida com a que é feita no mercado de trabalho (ver Lam, 1988). Assim, estas variáveis poderiam ser usadas como *proxies* de atributos não-observáveis das pessoas.

Do nosso ponto de vista, contudo, as variáveis utilizadas no presente artigo satisfazem, de forma razoável, a tarefa de retratar as condições socioeconômicas da família. Além disso, a utilização das escolaridades do cônjuge e dos sogros como *proxies*, sob a argumentação de que as "avaliações" feitas pelos "mercados" matrimonial e de trabalho em relação as características produtivas da pessoa são parecidas, é discutível. É forte a possibilidade de determinação no sentido inverso: o nível de rendimento da pessoa é que condiciona a escolaridade do cônjuge e dos sogros.

4.3 Estimando por MMQ e MVI

Quando comparamos a PNAD-96 com outras bases empregadas nos trabalhos internacionais, somos forçados a admitir que as informações disponíveis são ainda insuficientes para a análise pretendida aqui. Mesmo porque estes trabalhos normalmente utilizam *surveys* específicos (ou diversas pesquisas combinadas), cujos objetivos geralmente são mais limitados e bem mais dirigidos para o tema da educação. Assim, pelo fato de o escopo de pesquisa da

PNAD ser mais amplo e não tão específico, alguns dados utilizados nestes trabalhos não estão disponíveis para o nosso caso. Isto nos coloca algumas restrições. Em primeiro lugar, não temos condições de utilizar a abordagem que busca construir estimadores intrafamiliares, já que não existem dados adequados para isso nas PNADs. Na verdade, em princípio só seria possível fazer comparações entre pais e filhos, já que dispomos de informações sobre as escolaridades de ambos e sobre suas ocupações. Entretanto, não temos dados sobre os rendimentos dos pais, o que inviabiliza qualquer tentativa. Em segundo lugar, se formos aplicar o MVI, temos um certa dificuldade em encontrar instrumentos que estejam **fortemente** correlacionados com a educação, mas que não afetem diretamente os rendimentos (ver equações (8), (9) e (13)), já que se uma destas condições não for obedecida, os estimadores obtidos pelo MVI podem ser ainda mais inconsistentes do que os do MMQ. De início, os únicos candidatos disponíveis são algumas características das condições socioeconômicas da família, como a escolaridade dos pais ou o trabalho que ocupavam quando seus filhos tinham idade escolar, à semelhança dos empregados em diversos estudos (ver, por exemplo, Levin e Plug, 1999). Entretanto, encontramos evidências neste trabalho de que estas variáveis não satisfazem as condições para serem utilizadas como instrumentos, como já tínhamos argumentado na seção 2.1. De qualquer modo, será aplicado o MVI utilizando estas variáveis e seguindo a metodologia proposta por algumas pesquisas. Posteriormente, discute-se em que medida as estimativas obtidas são mais robustas do que as do MMQ.

Cabe dizer que, analisando os dois métodos apresentados, a abordagem que busca construir estimadores intrafamiliares com amostras de gêmeos univitelinos, quando combinado com o MVI, seria idealmente a mais indicada para estimar o efeito da educação sobre os rendimentos no Brasil. Ela não só dá uma solução superior em relação ao viés provocado pela omissão da variável genérica habilidade, como consegue isolar a influência do ambiente familiar sobre o rendimento individual e o grau de escolaridade atingida. Desta forma, não é preciso empregar *proxies* que mensurem imperfeitamente estes fatores, o que pode afetar a estimação. Contudo, não é demais dizer que neste caso a base de dados necessária é ainda mais difícil de ser obtida e trabalhada. Não só a amostra tende a ter um número bastante reduzido de observações, como os gêmeos utilizados precisam já ter completado seus estudos e estarem trabalhando. Ademais, apesar de terem crescido juntos no mesmo ambiente social e familiar, devem ter escolaridades diferentes. No Brasil, infelizmente ainda não há nenhuma iniciativa que busque construir um acervo de informações deste tipo.

Em relação ao viés provocado pela omissão de determinadas habilidades individuais na equação de rendimentos, foi visto que a ausência de uma *proxy* que mensure adequadamente estes fatores leva à superestimação do efeito da educação sobre os rendimentos. Como não contamos com informações sobre testes de QI ou semelhantes, não temos condições também de investigar este ponto de forma apropriada. Entretanto, acreditamos que a inclusão dos

fatores associados às condições socioeconômicas da família no modelo pode atenuar, em certa medida, o viés nas taxas de retorno da educação causado pela omissão destas variáveis. Isto porque, como no caso das características de personalidade, estas variáveis são, de certo modo, condicionadas pelo ambiente familiar e por suas origens paternas. Assim, os fatores associados às condições socioeconômicas da família podem captar parte da influência direta das habilidades individuais sobre os rendimentos. Por outro lado, reconhecemos aqui que o problema é um pouco diferente do caso das características de personalidade porque existem formas mais "adequadas" de mensuração destas habilidades, inclusive sugeridas na literatura especializada.

Desta forma, dentro de certas limitações, tentaremos apurar as nossas estimativas, procurando tratar dos três problemas discutidos: a omissão de variáveis relevantes, a endogeneidade da educação e possíveis erros de medida nesta variável. Ainda que os resultados obtidos estejam longe do ideal, eles fornecem elementos que permitem uma melhor compreensão da funcionalidade que a educação joga dentro da determinação dos rendimentos individuais e de temas correlatos, como o processo de distribuição de renda. Existem evidências, por exemplo, de que a taxa de retorno da escolaridade que geralmente é obtida pode estar superestimada. Mesmo assim, não há como negar a robusta evidência de que a educação é um dos fatores mais importantes para explicar a conformação das rendas de trabalho.

Três tipos de estimações são feitas: (i) uma nos moldes tradicionais (análoga às feitas usualmente na literatura nacional),[13] com uma equação sem as variáveis que retratam as condições socioeconômicas da família; (ii) outra incluindo estes fatores e verificando se eles são significativos para explicar os rendimentos individuais; (iii) e uma última aplicando o MVI e utilizando como instrumentos as variáveis associadas às condições socioeconômicas da família. Cabe dizer que somente no último caso não será empregado o modelo em que a educação é apresentada por variáveis binárias porque ele não apresentou um bom ajustamento, inclusive com taxas de retorno da educação negativas para o colegial.

Em primeiro lugar, estimamos os modelos (1) e (26) pelo MMQ, tendo por base uma equação sem as variáveis associadas às condições socioeconômicas da família (coluna (a) da Tabela 3) e outra com estas presentes (coluna (b) da Tabela 3). Os resultados são apresentados nas Tabelas 3 e 4.

13 Ver, por exemplo, Ramos e Vieira (1996).

Tabela 3
Estimativas de Dois Modelos de Equações de Rendimento Incluindo (*b*) ou Não (*a*) as Características dos Pais

Equação (1)			Equação (26)		
Variáveis	Coefic. para modelo		Variáveis	Coefic. para modelo	
	(a)	(b)		(a)	(b)
Constante	1,6931	1,6605	Constante	1,8495	1,7724
Homens	0,4451	0,4415	Homens	0,4506	0,4438
ID=Idade/10	0,5189	0,5546	ID=Idade/10	0,5009	0,5415
$(ID)^2$	–0,0507	–0,0544	$(ID)^2$	–0,0511	–0,0543
Escolaridade	**0,1129**	**0,0935**	**Escolaridade Superior**	**1,6625**	**1,3685**
			Colegial	**0,9538**	**0,7857**
			Ginásio	**0,5254**	**0,4335**
			Primário	**0,2684**	**0,2304**
Cor: Branca	0,1569	0,1281	Cor: Branca	0,1720	0,1385
Amarela	0,3694	0,3369	Amarela	0,3736	0,3397
Indígena	–0,0433 *ns*	–0,0416 *ns*	Indígena	–0,0127 *ns*	–0,0126 *ns*
Parda	0,0156 *ns*	0,0108 *ns*	Parda	0,0207 *ns*	0,0142 *ns*
Área Urbana	0,1722	0,1612	Área Urbana	0,1894	0,1732
Setor: Indústria	0,3336	0,3357	Setor: Indústria	0,3860	0,3700
Serviços	0,3150	0,3039	Serviços	0,3838	0,3526
Posição: Empregados	0,1656	0,1554	Posição: Empregados	0,2077	0,1882
Func. Públ.	0,2997	0,2959	Func. Públ.	0,3278	0,3172
Conta-Própria	0,2256	0,2105	Conta-Própria	0,2704	0,2440
Empregadores	0,9346	0,8825	Empregadores	0,9957	0,9319
Escol. Pai: Nível 5		0,3105	Escol. Pai: Nível 5		0,2861
Nível 4		0,1839	Nível 4		0,1843
Nível 3		0,1048	Nível 3		0,1269
Nível 2		0,0467	Nível 2		0,0701
Escol. Mãe: Nível 5		0,2903	Escol. Mãe: Nível 5		0,2436
Nível 4		0,2497	Nível 4		0,2337
Nível 3		0,1743	Nível 3		0,1817
Nível 2		0,0630	Nível 2		0,0838
Grupos Ocup.: Nível 6		0,2139	Grupos Ocup.: Nível 6		0,2124
Nível 5		0,2071	Nível 5		0,2173
Nível 4		0,1098	Nível 4		0,1159
Nível 3		0,0859	Nível 3		0,1061
Nível 2		0,0293	Nível 2		0,0535
Horas Trab.: de 15 a 39	0,4209	0,4262	Horas Trab.: de 15 a 39	0,4290	0,4325
de 40 a 44	0,7409	0,7460	de 40 a 44	0,7527	0,7554
de 45 a 48	0,7059	0,7189	de 45 a 48	0,7209	0,7305
49 ou mais	0,8679	0,8738	49 ou mais	0,8819	0,8851
Pessoa de Ref.	0,1212	0,1179	Pessoa de Ref.	0,1140	0,1139
Região: Norte	0,2509	0,2465	Região: Norte	0,2790	0,2660
RJ+ES+MG	0,2595	0,2473	RJ+ES+MG	0,2927	0,2674
SP	0,5725	0,5898	SP	0,6027	0,6073
Sul	0,2854	0,2940	Sul	0,3155	0,3084
Centro-Oeste	0,3352	0,3367	Centro-Oeste	0,3612	0,3536
n	54,757	54,757	*n*	54,757	54,757
Coef. Determin. Ajustado	0,5839	0,5965	*Coef. Determin. Ajustado*	0,5831	0,5950

Obs.: a notação *ns* significa que os coeficientes não são estatisticamente diferentes de zero ao nível de 5%.

O que observamos é que quando incluímos as variáveis associadas às condições socioeconômicas da família, a taxa de retorno da educação cai 18% na equação (1) - de 12,0% para 9,8% - e mais ainda no modelo (26) - as taxas de retorno para o primário, ginásio, colegial e superior, que eram de 11,3%, 6,6%, 13,0% e 17,1%, respectivamente, passam para 9,7%, 5,2%, 10,6% e 13,8%, com reduções de 14,2% a 21,2% -, indicando em que medida estes coeficiente estavam viesados. Ao realizar os testes usuais para verificar a significância dos coeficientes de regressão das variáveis associadas às condições socioeconômicas da família, todos apontaram que eles são estatisticamente diferentes de zero ao nível de significância de 1%, fornecendo-nos indicações que elas foram corretamente incluídas nos modelos da coluna (b), já que exercem uma influência direta sobre os rendimentos. Desta forma, podemos afirmar que a tendenciosidade provocada pela omissão dos fatores associados às condições socioeconômicas na equação de rendimentos é de considerável magnitude. O artigo de Lam e Schoeni (1993) já mostrava que a inclusão da escolaridade dos pais da pessoa reduz a estimativa da taxa de retorno da educação em cerca de 14%. Essa redução chega a superar 34% quando são incluídas as escolaridades da esposa e dos sogros. Cabe ressaltar que tanto a especificação do modelo da equação de rendimento como a delimitação da amostra são diferentes daquelas aqui utilizadas.[14]

Um outro resultado que reforça a conclusão do parágrafo anterior é que as contribuições marginais da educação diminuem cerca de 50% nas estimações dos dois modelos, o que significa que houve uma redução relevante no poder de explicação da escolaridade da pessoa quando as características dos pais são incluídas na equação de regressão. Ademais, mesmo descontados os efeitos da educação, ainda sobra uma pequena, mas significativa, contribuição marginal das três variáveis que mensuram as condições socioeconômicas da família, o que indica novamente que elas são relevantes na conformação dos rendimentos individuais.

Tabela 4
Contribuição Marginal (%) de Alguns Fatores para a Soma de Quadrados da Regressão

Fator	Equação (1)		Equação (26)	
	Contrib. Marginal (%) para o modelo		Contrib. Marginal (%) para o modelo	
	(*a*)	(*b*)	(*a*)	(*b*)
Escolaridade	24,19	11,64	24,09	11,43
Esc.Pai		0,22		0,22
Esc.Mãe		0,30		0,30
Grupos Ocup.Pai		0,30		0,34

14 Foi utilizada um amostra de homens na idade de 30 a 55 anos, com rendimentos positivo, da PNAD de 1982.

É possível que a influência das condições socioeconômicas da família esteja superestimada pelos motivos que discutimos anteriormente: ela pode estar captando os efeitos das habilidades individuais e das características de personalidade que afetam diretamente os rendimentos. Entretanto, verifica-se que os resultados são bastante significativos, indicando a existência de um considerável efeito da escolaridade e do *status* socioeconômico dos pais sobre o rendimento dos filhos. Podemos argumentar que mesmo levando em conta que a correlação entre os fatores associados às condições socioeconômicas e estes atributos seja forte, não é perfeita e, portanto, eles captariam apenas **alguma** influência dos últimos; o restante seria decorrente da existência de um efeito direto. Por outro lado, podemos acrescentar que as variáveis utilizadas para retratar as condições socioeconômicas da família - escolaridades dos pais e grupo de ocupação do pai - são insuficientes para mensurá-las. Como já explicamos, uma forma mais apropriada de se fazer isto seria empregar adicionalmente outras variáveis, como a renda dos pais e o grau de riqueza familiar. Entretanto, não dispomos destas informações. Nestes termos, indicamos que a influência do ambiente socioeconômico da família sobre os rendimentos deve estar subestimada. Vale dizer que a magnitude deste viés dependerá da correlação existente entre as variáveis empregadas e as outras não disponíveis. Se esta for forte, a omissão da renda dos pais ou da riqueza familiar terá importância reduzida.

Passemos agora a estimar o modelo (1) pelo MVI, para tratar dos possíveis problemas de erros de medida na educação e da endogeneidade desta variável. Como já explicamos, em princípio os únicos candidatos disponíveis para serem utilizados como instrumentos são os fatores associados às condições socioeconômicas da família. Entretanto, pudemos verificar, pelas estimações anteriores, que eles não satisfazem a condição expressa na equação (9) - de que eles não afetem diretamente os rendimentos -, o que invalidaria o emprego deles. Mesmo assim, diversos trabalhos acabam utilizando-os e tal procedimento parece útil para estabelecer um ponto de debate e apresentar a metodologia que geralmente é utilizada nestas análises. Contudo, fica a ressalva de que devemos olhar com desconfiança para os resultados apresentados.

Estimamos quatro equações de rendimento pelo método de mínimos quadrados em dois estágios, utilizando para as três primeiras um instrumento por vez (as variáveis binárias referentes a uma das características do pai ou da mãe) e, para a última, todos os fatores associados às condições socioeconômicas da família conjuntamente. Os resultados aparecem na Tabela 5.

Tabela 5
Estimação por MVI

Variáveis		Coeficientes conforme variáveis instrumentais utilizadas			
		VI esc.pai	VI esc.mãe	VI grupos ocup.	VI todas
Constante		1,6867	1,6938	1,6721	1,6691
ID=Idade/10		0,4965	0,4932	0,5053	0,4986
$(ID)^2$		–0,0452	–0,0446	–0,0463	–0,0447
Escolaridade		***0,1548***	***0,1586***	***0,1487***	***0,1611***
Cor	Branca	0,0759	0,0725	0,0858	0,0718
	Amarela	0,2070	0,1963	0,2263	0,1980
	Indígena	–0,0793 *ns*	–0,0818 *ns*	–0,0785 *ns*	–0,0806 *ns*
	Parda	0,0105 *ns*	0,0100 *ns*	0,0138 *ns*	0,0107 *ns*
Área	Urbana	0,1086	0,1067	0,1106	0,0995
Setor	Indústria	0,2820	0,2788	0,2829	0,2617
	Serviços	0,1456	0,1388	0,1525	0,1207
Posição	Empregados	0,1083	0,1004	0,1233	0,0891
	Func. Públ.	0,0901	0,0741	0,1200	0,0561
	Conta-Própria	0,1956	0,1894	0,2082	0,1778
	Empregadores	0,7682	0,7551	0,7975	0,7472
Escol. Pai	Nível 5		0,1367	0,1862	
	Nível 4		0,0216 *ns*	0,0427 *ns*	
	Nível 3		–0,0237 *ns*	–0,0123 *ns*	
	Nível 2		–0,0353	–0,0190 *ns*	
Escol. Mãe	Nível 5	0,0999		0,0757	
	Nível 4	0,0740		0,0742	
	Nível 3	0,0240 *ns*		0,0401	
	Nível 2	–0,0264		–0,0107 *ns*	
Grupos Ocup. Pai	Nível 6	0,1151	0,0505 *ns*		
	Nível 5	0,0620	0,0443		
	Nível 4	–0,0021 *ns*	–0,0123 *ns*		
	Nível 3	–0,0326	–0,0367		
	Nível 2	–0,0541	–0,0560		
Horas Trab.	de 15 a 39	0,4181	0,4159	0,4182	0,4133
	de 40 a 44	0,7725	0,7702	0,7722	0,7664
	de 45 a 48	0,7891	0,7881	0,7866	0,7854
	49 ou mais	0,9434	0,9423	0,9430	0,9403
Pessoa de Ref.		0,4593	0,4600	0,4559	0,4615
Região	Norte	0,2361	0,2379	0,2438	0,2355
	RJ+ES+MG	0,2292	0,2280	0,2319	0,2191
	SP	0,5544	0,5506	0,5567	0,5396
	Sul	0,2739	0,2737	0,2812	0,2617
	Centro-Oeste	0,3089	0,3087	0,3168	0,3042
n		54,757	54,757	54,757	54,757
Teste de Bassman		12,45 ***	6,31 ***	14,55 ***	19,29 ***
Teste de Hausman		137,28 ***	214,97 ***	181,08 ***	1438,73 ***
Coef. Determin. Ajustado		0,5346	0,5326	0,5377	0,5293

Obs: a notação *ns* assinala os coeficientes que não são estatisticamente diferentes de zero ao nível de significância de 5%. A notação *** indica que os testes são significativos a 1%.

De início, três coisas devem ser analisadas, conforme proposto por Bound *et al.* (1995): a qualidade das variáveis instrumentais utilizadas, se os problemas ligados aos erros de medida na educação e à endogeneidade desta variável são significativos e a validade dos instrumentos. Para verificar a qualidade dos instrumentos, como discutimos na seção 3.1, devemos testar se eles são fortemente correlacionados com a educação, já que, em caso contrário, apenas uma pequena associação entre o termo aleatório da equação de rendimento e a variável instrumental pode levar a uma grande inconsistência dos estimadores. Nestes termos, aplicamos o teste F no primeiro estágio da estimação para investigar se, conjuntamente, os coeficientes associados aos instrumentos excluídos da equação de rendimento são estatisticamente diferentes de zero (quer dizer, verificar se são relevantes para explicar o grau de escolaridade atingido por uma pessoa). Nas quatro equações, os valores obtidos foram bem altos, mostrando que eles são significativos ao nível de 1%, o que nos leva a concluir que os instrumentos propostos são muito bons tendo em vista esta condição: exercem uma influência direta sobre a educação e são fortemente correlacionados com ela.

A segunda questão pode ser analisada testando se os coeficientes obtidos pelo MVI são estatisticamente diferentes dos obtidos pelo MMQ. Se não forem, os problemas provocados pelos erros de medida na educação e pela endogeneidade desta variável não serão significativos, não justificando o emprego do MVI para corrigi-los. Isto pode ser feito pelo teste de Hausman, cujos resultados são apresentados na Tabela 5. Podemos verificar também que, nas quatro equações os coeficientes de regressão obtidos são estatisticamente diferentes dos estimados pelo MMQ, mostrando que os dois problemas citados são relevantes e, conseqüentemente, o uso de variáveis instrumentais é recomendado.

Por último, para a validação das variáveis instrumentais, deve-se investigar se os instrumentos não exercem uma influência direta sobre os rendimentos (que dizer, se eles não são correlacionados com o termo aleatório da equação de rendimento). Resultados anteriores já indicaram que no caso dos dados analisados ocorre aquela influência direta. Costuma-se verificar esta hipótese por meio de testes de superidentificação (teste de Bassman). A idéia é relativamente simples: caso o valor do teste dê muito alto (significativo), concluímos que existem evidências de que as variáveis exógenas do modelo foram indevidamente excluídas da equação de rendimento (no caso, as variáveis associadas às condições socioeconômicas da família). Podemos observar na Tabela 5 que em todas as equações o teste de Bassman deu significativo, fornecendo-nos evidências de que os instrumentos utilizados não deveriam ter sido excluídos das equações, já que exercem uma influência direta sobre os rendimentos. Assim, estes resultados confirmam o que já tínhamos apontado. Neste sentido, não seria apropriado estimar as equações pelo MVI utilizando os fatores associados às condições socioeconômicas da família como variáveis instrumentais, já que os estimadores dos parâmetros podem ser mais

inconsistentes do que os do MMQ. Desta forma, somos forçados a admitir que as estimativas apresentadas são pouco confiáveis e, portanto, devem ser vistas com reservas.

Para analisar os resultados propriamente ditos, seria mais correto, em primeiro lugar, comparar as três primeiras colunas da Tabela 5 com a equação (1) - coluna (*b*) da Tabela 3, onde a hipótese de linearidade entre os anos de escolaridade e o logaritmo do rendimento está presente e admitirmos que as condições socioeconômicas da família afetam os rendimentos em ambos os casos. O que notamos é que na Tabela 5 as taxas de retorno da escolaridade ficam aproximadamente entre 16,0% e 17,2%, ou seja, são 63% a 76% maiores do que o valor anteriormente estimado, de 9,8%. Comparando também a última coluna com a coluna (*a*) da Tabela 3 (onde as condições socioeconômicas da família não estariam especificadas no modelo), podemos verificar que a taxa de retorno obtida para a educação é 46% maior. Desta forma, podemos apontar que nos dois casos o viés associado com os problemas de endogeneidade da educação e de erros de medida pode ser relativamente grande, mas consistente com a magnitude encontrada em outros trabalhos (ver, por exemplo, Ashenfelter e Rouse, 1995 e Card, 1993). Cabe citar também que há uma certa estabilidade para as estimativas obtidas, mesmo utilizando diferentes instrumentos, o que é algo positivo. Entretanto, reiteramos uma vez mais que estes resultados podem estar bastante distorcidos em virtude do fato de as variáveis usadas como instrumentais terem efeito direto sobre os rendimentos.

Em suma, podemos apontar que os problemas referentes à correlação do termo aleatório da equação de rendimentos com a variável educação e/ou à omissão dos fatores associados às condições socioeconômicas da família são significativos na amostra, o que introduz um viés de considerável magnitude nas estimativas da taxa de retorno da escolaridade. Os resultados obtidos indicam que a omissão daquelas variáveis socioeconômicas causa um viés positivo próximo dos 20%. Por outro lado, não temos segurança para indicar a grandeza do viés negativo causado pelos demais problemas mencionados. Usando variáveis referentes às condições socioeconômicas dos pais como variáveis instrumentais, a taxa de retorno da educação estimada tornou-se cerca de 46% a 76% maior, mais do que compensando a redução na estimativa da taxa de retorno causada pela inclusão dessas variáveis. Sabemos, entretanto, que aquelas variáveis não têm as propriedades essenciais de uma boa variável instrumental.

5 Conclusões

Vimos que as mudanças na especificação da equação de rendimentos e no método de estimação podem ser extremamente importantes para avaliar a qualidade das estimativas do

retorno da educação no Brasil. Entretanto, uma limitação ainda a ser superada para poder aplicá-los refere-se à falta de dados. É de amplo conhecimento a excelência e a importância das informações coletadas pela PNAD no contexto brasileiro, mas em algum momento talvez seria relevante complementar o suplemento de mobilidade social com outras variáveis como a renda dos pais, ou efetuar uma pesquisa específica, dada a importância do tema.

É evidente que os problemas associados à aplicação do MMQ não se restringem apenas aos três discutidos. Pode-se argumentar, por exemplo, que a experiência profissional também sofre do problema de endogeneidade ou que é importante especificar no modelo a qualidade do ensino recebido.

Não existe ainda consenso de que os fatores associados às condições socioeconômicas da família afetam diretamente os rendimentos individuais, ainda que isto não seja verdadeiro para as habilidades individuais. Contudo, foi possível verificar que este não é o caso do Brasil, como antes já tínhamos argumentado. A omissão destes fatores no modelo pode introduzir um significativo viés positivo nas estimativas referentes à educação, ficando mais claro este fato no caso de especificarmos o modelo sem pressupor a existência de uma relação linear entre a educação e o logaritmo neperiano do rendimento. Nota-se que no País ainda são poucas as iniciativas de introduzir estes fatores nas análises econométricas dos efeitos da educação sobre os rendimentos individuais, cabendo destacar o artigo de Lam e Schoeni (1993). Utilizando uma PNAD anterior, ele encontrou resultados semelhantes para 1982, indicando a **persistência** da importância dos fatores associados às condições socioeconômicas na conformação dos rendimentos individuais no Brasil. Entretanto, vale argumentar que a inclusão de variáveis associadas às condições socioeconômicas no modelo pode agravar os problemas causados por erros de medida, levando a subestimar a taxa de retorno da educação.

Por outro lado, como foi visto neste artigo, é muito difícil mensurar a tendenciosidade provocada pelos erros de medida ou pela endogeneidade da escolaridade (quando o termo aleatório da equação de rendimento é correlacionado com a educação). As estimativas obtidas pelo MVI sugerem um viés negativo bem superior ao viés positivo causado pela omissão das variáveis associadas às condições socioeconômicas da família. Mas não há como negar a má qualidade dos instrumentos empregados, o que pode distorcer em muito os resultados obtidos. Pode acontecer até que os estimadores obtidos pelo MVI sejam mais inconsistentes do que aqueles obtidos pelo MMQ. Cabe ressaltar que a literatura internacional não tem dado a atenção devida a este fato. Assim, temos mais segurança em apontar apenas a tendenciosidade provocada pela omissão dos fatores associados às condições socioeconômicas da família.

Desta forma, podemos observar que a tarefa de estimar a taxa de retorno da escolaridade é mais complexa do que parece, o que exige cuidado nas pesquisas que buscam compreender

a funcionalidade da educação em temas correlatos como o da distribuição da renda e do combate à pobreza. Neste trabalho, encontramos uma taxa de retorno da educação mais próxima dos 9,8% do que dos 12,0%, e se quebrarmos a hipótese da relação linear entre a educação e o logaritmo neperiano do rendimento, a taxa para o primário, ginásio, colegial e nível superior fica mais perto dos 9,7%, 5,2%, 10,6% e 13,8%, respectivamente, do que dos 11,3%, 6,6%, 13,0% e 17,1%, respectivamente. Cabe enfatizar que estas estimativas podem estar bastante viesadas devido aos erros de medida na educação e à endogeneidade desta variável.

Bibliografia

Ashenfelter, O., Rouse, C. *Schooling, intelligence and income in America*: cracks in the bell curve. Cambridge, MA. National Bureau of Economic Research (Nber), 1995 (Working Paper Series, n. 6902).

Becker, G. *Human capital*: a theorical and empirical analysis, with special reference to education. Nova York: Columbia University Press, 1964.

Bound, J., Jaeger, D.; Baker, R. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variables is weak. *Journal of the American Statistical Association*, v. 90, p. 443-450, 1995.

Bowles, S. Schooling and inequality from generation to generation. *Journal of Political Economy*, v. 80, 1972.

_____ Understanding unequal economic opportunity. *American Economics Review*, v. LXIII, n. 2, 1973.

Card, D. *Using geographic variation in college proximity to estimate the return to schooling*. Cambridge, MA. National Bureau of Economic Research (Nber), 1993 (Working Paper Series, n. 4483).

_____ The causal effects of education on earnings. *In*: Ashenfelter, O.; Card, D. (orgs.), *Handbook of Labor Economics*, v. 3, 1999.

Griliches, Z. Estimating the returns to schooling: some ecomometric problems. *Econometrica*, v. 45, n. 1, p. 1-22, 1977

Hoffmann, R. Mensuração da desigualdade e pobreza no Brasil. *In*: Henriques, R., *Desigualdade e pobreza no Brasil.* Rio de Janeiro: IPEA, 2000.

Isacsson, G. Estimates of the return to schooling in Sweden from a large sample of twins. *Labor Economics*, n. 6, p. 471-489, 1999.

Jencks, C. *A Reassessment of the effect of family and schooling in America*. Nova York: Basic Book, 1972.

Lam, D. Marriage markets and assortative mating with household public goods: theorical results and empirical implications. *J. Human Resources*, n. 23, p. 462-87, Fall 1988.

Lam, D., Schoeni, R. Effects of family background on earnings and returns to schooling: evidence from Brazil. *Journal of Political Economy*, v. 101, n. 4, p. 710-740, 1993.

Levin, J., Plug, E. Instrumenting education and the returns to schooling in the Netherlands. *Labor Economics*, n. 6, p. 521-534, 1999.

Mincer, J. *Schooling, experience and earnings*. Nova York: National Bureau of Economic Research, 1974.

Pastore, J., Silva, N. *Mobilidade social no Brasil*. São Paulo: Makron, 2000.

Ramos, L., Vieira, M. A relação entre educação e salários no Brasil. *In*: *A economia brasileira em perspectiva*. Rio de Janeiro: IPEA, 1996.

Savedoff, W. Os diferenciais de regionais de salários no Brasil: segmentação versus dinamismo da demanda. *Pesquisa e Planejamento Econômico*, v. 20, n. 3, p. 521-555, 1990.

Ueda, E. *Educação e rendimento: uma abordagem econométrica.* 2001. Dissertação (Mestrado), Unicamp/IE, Campinas.

Willis, R. Wage determinants: survey and reinterpretation of human capital earnings function. *In*: Ashenfelter, O., Layard, R (orgs.), *Handbook of labor economics,* v. 3, 1986.