3
Estratégia Empírica

Na busca pelos efeitos da educação dos pais na educação dos filhos, utilizamos duas estratégias. Primeiro, estimamos esse efeito através do método de Mínimos Quadrados Ordinários, inserindo progressivamente controles que isolam os efeitos procurados de outros fatores não-observados que estejam influenciando a relação estimada. Em uma seção seguinte, procuramos estimar o real efeito causal da educação dos pais nos filhos, através do método de Mínimos Quadrados em Dois Estágios (2SLS) com variáveis instrumentais. Para tal, devemos ter claro o modelo gerador de educação dos filhos, as limitações a que estamos sujeitos nessa análise empírica, e as potenciais formas de superá-las.

3.1.
O Modelo

Utilizamos um modelo de transmissão intergeracional de educação *à la* Becker e Tomes (1986), com algumas peculiaridades. O capital humano de uma família se relaciona da seguinte forma:

$$ed^i_{filho} = \beta * ed^i_{pais} + \delta * X^i + u^i_{filho} \qquad (1)$$

Onde ed^i_{filho} e ed^i_{pais} são, respectivamente, a educação de um indivíduo da geração da família *i* e seus pais, X é um vetor de características observáveis de filho, mãe e pai, e u^i_{filho} é um erro aleatório não-observável.

Portanto, o capital humano do filho depende de fatores observáveis, como sua idade e a renda de seus pais, mas também diretamente do capital humano de pai e de mãe. É o caso de complementaridade entre educação dos pais e dos filhos na função de produção de educação da criança. Nesse sentido é que chamamos β o efeito causal da educação de uma geração sobre a outra.

O erro *u* contém fatores que influenciam o nível de capital humano do filho. Como enfatiza a literatura, há um conjunto enorme de fatores que podem estar exercendo esse papel. Basicamente, podemos dividi-los em três grupos: (i)

fatores familiares, passados de geração a geração, como cultura familiar, preferência por trabalho e/ou estudo, talento e habilidade ou inteligência; (ii) fatores comuns a todos em um dado período de tempo ou certo lugar, (iii) atributos exclusivamente pessoais do indivíduo.

Matematicamente:

$$u^i_{filho} = v^i_{filho} + w^i + \varepsilon^i_{filho} \qquad (2)$$

Em que v^i_{filho}, w^i e ε^i_{filho} correspondem respectivamente a (i), (ii) e (iii) do parágrafo anterior. As hipóteses que se fazem são de que o termo ε^i_{filho} é um erro estocástico não-correlacionado com qualquer outra variável e de que os fatores familiares são transmitidos da seguinte forma:

$$v^i_{filho} = \rho * v^i_{pais}, \quad 0 < \rho < 1 \qquad (3)$$

A partir das três equações acima fica claro que $Cov(ed^i_{pais}, u^i_{filho}) \neq 0$. Assim, ao fazermos a projeção de ed^i_{filho} em ed^i_{pais}, estaremos captando não apenas o efeito causal deste naquele, mas uma correlação entre essas variáveis que incluirá (i) o efeito dos fatores familiares não-observáveis que influenciam educação de pais e filhos simultaneamente; (ii) os efeitos dos canais através dos quais quaisquer outras variáveis não incluídas entre os controles (como a renda dos pais, por exemplo) influenciem o capital humano do filho. É isso que motiva o estudo mais cuidadoso dessas relações, através tanto de modelos MQO mais bem especificados quanto da busca por variáveis instrumentais que sejam capazes de isolar esse efeito causal.

Como vamos distinguir entre os efeitos de mãe e pai, o modelo estimado é um pouco mais abrangente. As equações são as seguintes:

$$ed^i_{filho} = \alpha_j * ed^i_{mãe} + \beta_j * ed^i_{pai} + c * X^i + u^i_{filho}, \quad j = h, m \qquad (4)$$

$$u^i_{filho} = \lambda * u^i_{mãe} + \phi * u^i_{pai} + \varepsilon^i_{filho}, \quad 0 < \lambda < 1, \quad 0 < \phi < 1 \qquad (5)$$

Em caso de *matching* no mercado de casamentos, temos ainda:

$$Cov(u_{mae}, u_{pai}) \neq 0 \qquad (6)$$

Em que α_j e β_j são respectivamente o efeito da educação da mãe e do pai no filho de sexo *j*, o distúrbio *u* é influenciado tanto pelo distúrbio materno quanto paterno, e ε^i_{filho} é o distúrbio não-correlacionado.

3.2.
As Estimações por MQO

O primeiro conjunto de regressões busca identificar padrões na correlação entre os níveis de capital humano observados. Idealmente, gostaríamos de uma medida do capital humano final atingido pelo filho. No entanto, a PNAD não possui esses dados, uma vez que os filhos que têm seus dados coletados são aqueles que ainda vivem com os pais e, em sua grande maioria, têm menos que 18 anos[10]. A solução encontrada é seguir uma estratégia parecida com Horowitz e Souza (2004), criando uma medida de defasagem idade-série *normalizada* que, sob hipóteses bastante gerais, é uma *proxy* para o nível de capital humano atingido no futuro[11].

O Ministério da Educação (MEC) determina que toda criança com 7 anos completos até 1º de março deve estar matriculada na 1ª série. Se a criança satisfizer esse requisito e não for reprovada em nenhum ano, quando ela tiver 7+y anos completos até 1º de março ela estará na y+1-ésima série. A medida de defasagem idade-série *bruta* é simplesmente quantos anos de estudo completos a menos tem uma criança em relação àquela com desempenho ideal. Por ser uma variável limitada cujo conjunto de valores possíveis se desloca com o avanço da idade da criança, devemos fazer uma normalização para podermos utilizar o modelo clássico de regressão linear. Nossa variável dependente será:

$$y_i = \left(\frac{defas_i}{idade_i - 7} \right) x\, 100$$

Em que $defas_i$ é defasagem idade-série bruta do filho *'i'* e $idade_i$ é a idade escolar, ou seja, sua idade em 1º de março do ano em que a pesquisa foi feita. Um valor de y igual a zero representa um filho com percurso escolar ideal, sem

[10] Há um número considerável de filhos maiores de 25 anos que vivem com os pais. Porém, esta característica está possivelmente correlacionada com muitas outras que influenciam o nível de educação atingido, isto é, há um forte viés de seleção, o qual nos leva a desconsiderar a análise dos integrantes desse grupo.

[11] Diferentemente de Machado e Gonzaga (2007) e Oreopoulos *et al.* (2006) nossa medida de defasagem idade-série não é binária. Portanto, não indica a probabilidade de um aluno já ter sido reprovado em alguma série.

atrasos. Um valor igual a 100 representa um filho *totalmente atrasado*[12]. Desse modo, podemos pensar nossa variável como uma medida do percentual de atraso da criança em relação ao desempenho ideal.

Com essa variável estamos capturando o efeito da educação dos pais em (i) a probabilidade de o filho ser reprovado 'vezes' o número de anos que ele já está na escola, (ii) a chance de o filho largar a escola antes de completar 18 anos e (iii) o filho entrar atrasado na escola *desde o início*.

Apesar de estarmos englobando três efeitos diferentes em uma mesma variável, acreditamos que todos estarão correlacionados com o nível final de capital humano atingido pelos filhos, que é o que gostaríamos de avaliar. Além disso, dada a enorme porcentagem de alunos no Brasil que têm defasagem idade-série estritamente positiva, essa variável tem vantagens sobre aquela utilizada por Machado e Gonzaga (2007)[13], nos permitindo distinguir e quantificar os casos mais adequadamente, com também avaliar possíveis não-linearidades existentes.

Apesar da falta de modelos satisfatórios que impliquem diferentes efeitos tanto de pai *versus* mãe quanto sobre meninos *versus* meninas, uma crescente parte da literatura empírica tem abordado essa questão, com bons indícios da existência dessas diferenças. Portanto, a princípio estimaremos separadamente as relações mãe-filho, mãe-filha, pai-filho e pai-filha.

Nossa especificação básica inclui como controles a idade escolar, idade escolar ao quadrado, idade e idade ao quadrado de mãe e pai, além de *dummies* que controlam para efeitos fixos de: gênero e cor do filho e Unidade da Federação de nascimento de filhos, mães e pais.

Em seguida, incluímos entre os controles a renda per capita familiar para ver o quanto da transmissão intergeracional de educação estimada inicialmente está se dando *através* da renda.

É mais provável que o investimento em capital humano esteja sendo influenciado por uma variável de estoque de riqueza, e não de fluxo. Para tratar do caso em que a medida relevante seja a renda permanente, construímos variáveis indicadoras de características domiciliares e as incluímos numa 3ª especificação.

---

[12] Dessa forma, um filho com 11 anos de idade cuja variável dependente assume o valor 50 está 2 anos atrasado, enquanto um filho com 15 anos com o mesmo valor está 4 anos atrasado.

[13] Ver o Capítulo 2

Por último, adicionamos como controles variáveis que indicam o nível educacional atingido por avós paternos e maternos, que podem conter informação sobre (i) renda permanente familiar; (ii) características familiares não-observáveis. A essa especificação damos o nome "completa". Em todas as estimações permitimos *clustering* do erro padrão no nível da família.

Por mais que as equações estimadas sejam bem especificadas, a existência de fatores familiares não-observáveis faz com que as estimativas MQO sejam viesadas. De (1), temos:

$$ed^i_{filhos} = \beta * ed^i_{pais} + \delta * X^i + u^i_{filhos}$$
$$ed^i_{pais} = \beta * ed^i_{avós} + \delta * X^i + u^i_{pais}$$

Por simplicidade, vamos ignorar os X^i. Lembrando de (2) e (3), então:

$$\hat{\beta}_{MQO} = \beta + \rho * Cov(ed_{pais}, u_{pais}) \big/ Var(ed_{pais})$$

Logo, ainda que incluíssemos todas as características observáveis relevantes e perfeitamente medidas, o estimador de β por MQO ainda possuiria um viés positivo.

3.3. A Busca pelo Efeito Causal: a Estratégia de Variáveis Instrumentais

Como explicado acima, não se pode afirmar que os estimadores MQO encontrados sejam evidência de uma causalidade direta como aquela encontrada em Black *et al.* (2005) e Oreopoulos *et al.* (2006). Para isolar os efeitos de simultaneidade e hereditariedade existentes devido à correlação intergeracional de habilidade, talento, cultura ou outros, recorreremos à utilização de variáveis instrumentais.

Um bom instrumento tem que satisfazer dois requisitos: (i) ser não-correlacionado com o erro da equação estimada; (ii) ser suficientemente correlacionado com a variável que se supõe endógena. O segundo requisito é observável através de testes estatísticos simples; o primeiro tem como principal suporte a argumentação teórica.

O grupo de variáveis que utilizaremos como instrumentos já foi usado em Emerson e Souza (2007) e Machado e Gonzaga (2007). Trata-se do número de

escolas por habitante e do número de professores por escola na Unidade da Federação em que o pai (ou a mãe) nasceu, no ano de seu nascimento.

Sabemos que a oferta de serviços educacionais no Brasil durante o século XX foi bastante precária na maior parte do país. Supomos que variações nessas séries em certa UF em dado ano representam variações na oferta de educação para as crianças que estavam em idade escolar naquela UF àquela época. Essas variações na oferta educacional devem estar relacionadas a variações no custo e no benefício de se educar. Devem, portanto, estar correlacionadas com o nível de educação adquiridos por essas crianças.

Além disso, parece razoável supor que essas variações *não* estão correlacionadas com variações das características não-observáveis de cada família ou indivíduo que poderiam estar gerando os problemas de simultaneidade e viesando os coeficientes estimados. Se isso for verdade, então essas variáveis podem constituir bons instrumentos para a educação dos pais, e o método de 2SLS pode nos dar os reais coeficientes α_j e β_j da equação (4).

Embora a hipótese de que essas variações nas ofertas de escolas e de professores sejam exógenas aos níveis familiar e individual pareça coerente, devemos fazer uma análise cuidadosa para que nossos instrumentos estejam capturando o efeito desejado, que é o de variações exógenas na oferta de serviços escolares. Por se tratarem de séries históricas anuais, também poderiam estar indiretamente caracterizando melhorias ocorrendo em várias outras dimensões da provisão de bens públicos, possivelmente muito correlacionadas com as coortes de nascimentos dos indivíduos. Essa questão será tratada no capítulo 6.