UMA ANÁLISE ECONOMÉTRICA DOS DETERMINANTES DO ACESSO À UNIVERSIDADE DE SÃO PAULO*

Daulins Rêni Emilio
Da Boston University
Walter Belluzzo Jr.
Da USP
Denisard C. O. Alves
Da USP

O acesso ao ensino superior costuma ser visto como um fator fundamental para a redução da desigualdade de renda no Brasil. Este artigo apresenta uma análise dos determinantes do acesso à universidade pública, que é potencialmente útil para a análise de políticas direcionadas à ampliação das oportunidades de acesso ao ensino superior público. Utilizando dados do vestibular 2000 da Fundação Universitária para o Vestibular (Fuvest) e da Pesquisa Nacional por Amostra de Domicílios (PNAD) de 1999, foi estimado um modelo econométrico em que é corrigido o viés de seleção que surge como conseqüência do fato de que apenas os indivíduos que se candidataram são observados.

1 INTRODUÇÃO

A redução do nível de concentração de renda no Brasil é um dos principais desafios de política pública. Estudos recentes têm mostrado que a desigualdade de renda no Brasil está estreitamente relacionada às diferenças educacionais observadas entre indivíduos. Barros e Mendonça (1996), por exemplo, estimam que a eliminação dos diferenciais de renda por nível educacional poderia levar a uma redução entre a metade e 1/3 na desigualdade de renda.

Segundo Fernandes e Narita (2001), a influência marcante da educação sobre a desigualdade de renda no Brasil ocorre por dois motivos: a elevada desigualdade educacional entre os trabalhadores e a elevada sensibilidade dos salários em relação ao nível educacional. Consideradas as diferenças entre os vários níveis de educação e as diferentes carreiras, o rendimento salarial esperado por ano adicional de estudo superior sobressai dentre todos os demais níveis educacionais, ultrapassando a taxa de 20% [Fernandes e Menezez-Filho (2000)].

Nesse contexto, ao oferecer ensino gratuito a universidade pública assume um papel importante como instrumento de política social. Esse papel fundamental da universidade pública tem motivado a discussão de mecanismos para melhorar as condições de acesso para aqueles oriundos de classes menos favorecidas. De maneira geral, essa discussão tem sido direcionada para a reavaliação dos mecanismos de

* Este artigo tem como base a dissertação de mestrado apresentada ao Instituto de Pesquisas Econômicas, da FEA-USP, por Daulins Rêni Emilio. Os autores agradecem a Paulo Picchetti pelos comentários e pelas sugestões. Daulins Rêni Emilio agradece também ao CNPq pelo apoio financeiro.

financiamento público do ensino superior e dos critérios que determinam o acesso à universidade pública. Especificamente, são discutidas, atualmente, a cobrança de mensalidades nas universidades públicas, a criação de sistemas de cotas inspirados na "ação afirmativa" americana e a definição de notas de corte diferenciadas.

Para que sejam delineadas políticas coerentes de acesso ao ensino superior público, é fundamental que existam estudos detalhados sobre os determinantes do acesso à universidade pública, analisando o perfil daqueles que se beneficiam do financiamento público do ensino superior no Brasil. No entanto, não existem muitos trabalhos publicados analisando essa questão. Dentre os existentes, destacam-se as análises de Siano (1977), Freitas (1979) e Oliveira (1980).

A abordagem-padrão nesses trabalhos é a utilização do método de mínimos quadrados ordinários (MQO) e uma extensa análise descritiva dos dados disponíveis. O modelo considerado, geralmente, tem como variável dependente o número de pontos alcançados pelo candidato no vestibular e como variáveis explicativas a educação dos pais, a habilitação do corpo docente onde os candidatos realizaram seus estudos de segundo grau e o sexo do candidato.

Apesar de oferecerem elementos importantes para a elaboração de políticas de acesso, esses trabalhos não levam em consideração o fato de que as amostras disponíveis para esse tipo de estudo são truncadas por natureza — apenas aqueles inscritos no vestibular são observados. Desse modo, os resultados obtidos naqueles estudos referem-se aos determinantes do desempenho no vestibular condicional no fato de o candidato ter efetuado sua inscrição.

A hipótese implícita na utilização de MQOs é que a inscrição no vestibular é um processo aleatório e que não há correlação entre o processo de inscrição e o resultado final do vestibular. No entanto, o senso comum sugere que características observáveis e não-observáveis influenciam não só a probabilidade de o indivíduo obter sucesso no vestibular, como também a decisão anterior, de inscrever-se ou não para realizar o exame. Nesse caso, ao deixar de considerar o processo decisório da inscrição no vestibular ocorre o conhecido viés de seleção, como indicado por Heckman (1979).

O principal objetivo deste trabalho é propor um procedimento para corrigir o viés de seleção em estudos sobre o ingresso na universidade pública, aplicando-o ao caso da Universidade de São Paulo (USP). Especificamente, propomos um modelo de variáveis latentes para explicar o acesso à USP. A primeira variável latente leva a um modelo para explicar a inscrição, enquanto a segunda leva a um modelo para explicar o sucesso no vestibular. Como apenas indicadores discretos do nível dessas variáveis são observados, utilizamos modelos de escolha binária em ambos os casos. Desse modo, para corrigir o viés de seleção aplicamos o método proposto por Van de Ven e Van Praag (1981) para corrigir o viés de

seleção em um modelo *probit*. Esse método pode ser visto como uma adaptação de Heckman (1979), e tem sido utilizado na literatura em diversas circunstâncias. Painter (2000) e Painter, Gabriel e Myers (2001), por exemplo, aplicam esse modelo para explicar a escolha da propriedade de residência; Amuedo-Dorantes e Kimmel (2003) para a decisão de adiar a maternidade; e Ghidoni (2002) para a decisão dos jovens de sair do domicílio dos pais, dentre outros.

Os modelos foram estimados utilizando um conjunto de características que, presumivelmente, influenciam o acesso à universidade. Essas características incluem o tipo de ensino fundamental e médio, o tempo dedicado a cursos preparatórios, o grau de escolaridade dos pais e um conjunto de informações relacionadas às condições de renda domiciliar. Como essa amostra é truncada, para viabilizar a eliminação do viés de seleção é preciso obter informações também sobre os indivíduos que não se inscreveram no vestibular. Em outras palavras, é preciso uma amostra censurada. Neste artigo, propomos utilizar a PNAD de 1999 para extrair essas informações. Os dados disponíveis sobre os candidatos referem-se ao vestibular 2000 da Fuvest.[1]

Os resultados obtidos revelam que a correção do viés de seleção tem como efeito mais importante a redução dos efeitos marginais. Dentre as variáveis com maior efeito marginal sobre a probabilidade de sucesso no vestibular, destacam-se os cursos pré-vestibular, o nível superior da mãe, o ensino médio em período integral e em escolas federais ou estrangeiras. Finalmente, constatamos um efeito negativo, embora relativamente pequeno, relacionado às raças negra e parda.

O artigo está organizado da seguinte maneira: a Seção 2 apresenta o modelo econométrico a ser utilizado. Na Seção 3 discute-se o procedimento utilizado para consolidar as informações sobre os inscritos (Fuvest) e sobre os não-inscritos no vestibular (PNAD). A Seção 4 apresenta os dados utilizados. A Seção 5 mostra os resultados obtidos. As conclusões são apresentadas na Seção 6.

2 MODELO ECONOMÉTRICO

Como já discutido, o ponto central deste artigo é levar em consideração a decisão de participar do vestibular na análise dos determinantes do acesso à universidade pública. Para tanto, é preciso modelar tanto a participação quanto o sucesso no vestibular. Suponha que o sucesso no vestibular, entendido como o ingresso na universidade, seja determinado por uma variável latente que representa a habilidade do candidato. Admitindo que essa habilidade é determinada pelas características

1. Cabe destacar que as provas do vestibular 2000 foram aplicadas no final de 1999, para os ingressantes em 2000. Por isso utilizamos a PNAD de 1999.

socioeconômicas e pela formação familiar dos candidatos, podemos descrever que:

$$y_i^* = x_i'\beta + \varepsilon_i \tag{1}$$

em que x é um vetor de características individuais, β é um vetor de parâmetros e $\varepsilon \sim N(0, \sigma^2)$ é um termo aleatório refletindo efeitos não-capturados em x. Desse modo, o sucesso no vestibular pode ser descrito por uma variável binária, definida como:

$$S_i = \begin{cases} 1 & \text{se} \quad y^* \geq c_i \\ 0 & \text{caso contrário} \end{cases} \tag{2}$$

em que c_i reflete os diferentes graus de dificuldade associados a cada carreira.

Para completar o modelo, é preciso considerar que as notas são observáveis apenas para aqueles que decidiram pela participação no vestibular. Para tanto, é necessário definir, explicitamente, o processo de inscrição no vestibular. De certa maneira, a inscrição corresponde a uma demanda por ensino superior. Atualmente, a análise dessa demanda através de modelos teóricos é, relativamente, padrão na literatura, contando, inclusive, com um capítulo no *Handbook of Labor Economics* [Freeman (1986)]. Dentre os modelos existentes, destacam-se aqueles que se baseiam na teoria do capital humano [Becker (1964)] e na teoria da sinalização de mercado [Spence (1973)]. A teoria do capital humano vê a educação como um insumo produtivo, de modo que a demanda por educação é tratada do ponto de vista de um investimento. A teoria da sinalização de mercado considera que os indivíduos demandam educação como forma de sinalizar para o mercado de trabalho uma produtividade maior.

Uma vez que a identificação da demanda por educação não faz parte do escopo deste artigo, adotamos aqui uma abordagem generalista e apenas associamos a demanda por educação à "propensão ao ensino superior". Especificamente, admita que a decisão de participar do vestibular é determinada por uma variável latente que representa a propensão ao ingresso no ensino superior, I^*, definida de modo que o indivíduo i participe do vestibular apenas quando $I_i^* \geq 0$. Admita ainda que essa propensão pode ser representada por:

$$I_i^* = h_i'\gamma + \eta_i \tag{3}$$

em que h é um vetor de características do indivíduo, γ é um vetor de parâmetros e $\eta \sim N(0, \omega^2)$ é um termo aleatório refletindo efeitos não-capturados por h. Obviamente, observamos apenas uma variável binária,

$$I_i = \begin{cases} 1 & \text{se} \quad I^* \geq 0 \\ 0 & \text{caso contrário} \end{cases} \tag{4}$$

que indica o nível da variável latente, I^*.

Em conjunto, as equações (1) a (4) definem um modelo completo de acesso à universidade pública. Note-se que se habilidade fosse observável para todos os indivíduos em condições de ingressar no ensino superior, a equação (2) seria desnecessária e a estimação do modelo poderia ser implementada utilizando o método proposto por Heckman (1979). Como essa variável não é observável, uma alternativa nessa direção seria utilizar a nota obtida pelo candidato no vestibular como uma *proxy* para sua habilidade, de modo que c_i seria interpretado como a nota de corte para a carreira escolhida pelo candidato (isto é, a nota do último candidato convocado para a matrícula). No entanto, como as informações sobre as notas não foram disponibilizadas pela Fuvest, utilizaremos o modelo em que (2) é incluída.

A estimação do modelo definido pelas equações (1) a (4) exige uma adaptação do método de estimação proposto por Heckman (1979), pois este se aplica apenas ao caso em que a variável dependente é contínua. Especificamente, utilizaremos o modelo *probit* com seleção proposto por Van de Ven e Van Praag (1981). No entanto, para facilitar a exposição, é conveniente apresentar o caso de mínimos quadrados antes de introduzir o modelo *probit*. Como veremos a seguir, uma vez que o problema de mínimos quadrados é equacionado, a adaptação proposta por esses autores é direta.

Segundo Heckman (1979), aplicando mínimos quadrados diretamente sobre (1) estaríamos estimando, na verdade,

$$E(y^* \mid x, I^* \geq 0) = x_i' \beta + E(\varepsilon \mid x, I^* \geq 0) \tag{5}$$

Claramente, as estimativas obtidas dessa forma seriam adequadas apenas quando a esperança condicional de ε fosse igual a 0. Nos casos em que I^* não for independente de x, essa esperança condicional será diferente de 0, implicando que o estimador de mínimos quadrados será viesado. Especificamente, considerando que ε e η seguem uma distribuição normal bivariada padronizada com

coeficiente de correlação ρ e definindo $\lambda_i = \phi(z_i)/\Phi(-z_i)$, com $z_i = -h_i'\alpha/\omega$ e Φ representando a função distribuição normal, temos que:[2]

$$E(\varepsilon_i \mid x_i, I^* \geq 0) = \sigma\rho\lambda_i \tag{6}$$

de modo que os coeficientes estimados serão viesados se $\rho \neq 0$. Nesse caso, o modelo correto a ser estimado, normalizando $\sigma = 1$, é dado por:

$$y_i^* = x_i'\beta + \rho\lambda_i + v_i \tag{7}$$

em que $v = \varepsilon - \rho\lambda$. Em outras palavras, é preciso introduzir uma nova variável explicativa λ que, junto com o coeficiente associado a ela, assume o papel de um termo de correção que elimina o viés de seleção, já que $E(v \mid I^* \geq 0) = 0$. No entanto, é preciso notar que, apesar de eliminar o viés, a correção introduz heteroscedasticidade no modelo. Heckman (1979, p. 156-157) demonstra que:

$$\tau^2 = E(v^2 \mid I^* \geq 0) = 1 + \rho^2\lambda_i(z_i - \lambda_i) \tag{8}$$

Logo, para corrigir a ineficiência decorrente da heteroscedasticidade é preciso corrigir mais uma vez o modelo, utilizando (8).

O problema para implementar a correção indicada em (7) é que λ_i não é observável. Portanto, para implementar o modelo corrigido é preciso um estimador consistente de λ. O procedimento de estimação proposto pelo autor consiste em obter estimativas das probabilidades de cada indivíduo pertencer à amostra não-censurada através de um modelo *probit*. Utilizando essas probabilidades e os parâmetros estimados no *probit*, é possível calcular estimativas de λ e corrigir a regressão de interesse, considerando apenas as observações não-censuradas.

A aplicação do procedimento de Heckman a nosso caso exige alguma adaptação. Embora o princípio geral do termo de correção introduzido no modelo permaneça, é preciso considerar que o modelo a ser corrigido aqui é um *probit*, e não um modelo de regressão linear simples. Van de Ven e Van Praag (1981) sugerem dois procedimentos para implementar a correção à maneira de Heckman no modelo *probit*. O primeiro consiste em estimar conjuntamente todos os parâmetros de interesse por máxima verossimilhança.

Para simplificar a exposição sobre esses procedimentos de estimação, começamos ignorando o problema de seleção, isto é, consideramos apenas o modelo

2. A expressão λ é conhecida como Razão de Mills Invertida.

implicado por (2). Admitindo que ε tem distribuição normal, a estimação desse modelo através do método de máxima verossimilhança consiste na otimização de:

$$L = \prod_{i=1}^{N} [1 - \Phi(x_i'\beta)]^{1-S_i} \Phi(x_i'\beta)^{S_i} \tag{9}$$

em que Φ representa a função distribuição normal. A função de verossimilhança (9) é obtida considerando-se que cada S_i é uma realização de um evento de Bernoulli com probabilidade de sucesso ($S_i = 1$) igual a $\Phi(x_i'\beta)$. Desse modo, a probabilidade de se observar a seqüência de uns e zeros como aquela da amostra é dada pelo produtório em (9). Como há apenas dois eventos possíveis, $S = 1$ e $S = 0$, é conveniente reescrever a equação (9) fazendo referência explícita a esses eventos, isto é,

$$L = \prod_{S=1} \Phi(x_i'\beta) \prod_{S=0} \Phi(-x_i'\beta) \tag{10}$$

em que utilizamos o fato de Φ ser simétrica.

No caso do modelo completo, incluindo a equação de seleção, é possível obter a função de verossimilhança de maneira análoga. Primeiro, é preciso considerar que aqui existem três eventos possíveis: inscrito com sucesso, inscrito sem sucesso e não-inscrito. Além disso, uma vez que há duas variáveis latentes, é preciso considerar a distribuição conjunta dos erros. Sob a hipótese de que os erros seguem uma distribuição normal bivariada, as probabilidades associadas a cada evento são dadas por:

$$\begin{aligned} &I = 1, S = 1 : \text{Prob}\,(I = 1, S = 1) = \Phi_2(x_i'\beta, h_i'\alpha; \rho) \\ &I = 1, S = 0 : \text{Prob}\,(I = 1, S = 0) = \Phi_2(-x_i'\beta, h_i'\alpha; -\rho) \\ &I = 0 : \text{Prob}\,(I = 0) = \Phi(h_i'\alpha) \end{aligned} \tag{11}$$

em que Φ_2 representa a função distribuição normal bivariada [Greene (2000, p. 857)]. Dessa forma, segue-se que a função de verossimilhança para o modelo completo é dada por:

$$L = \prod_{\substack{I=1 \\ S=1}} \Phi_2(x_i'\beta, h_i'\alpha; \rho) \prod_{\substack{I=1 \\ S=0}} \Phi_2(-x_i'\beta, h_i'\alpha; -\rho) \prod_{I=0} \Phi(h_i'\alpha) \tag{12}$$

Em contraste com o modelo *probit* usual associado a (10), a otimização de (12) é computacionalmente pesada. Por isso, os modelos de seleção são, freqüentemente, estimados utilizando variantes do procedimento de dois estágios proposto por Heckman (1979) para evitar a otimização de funções de verossimilhança que envolvem a função distribuição normal bivariada, mas que leva a uma boa aproximação dos resultados que seriam obtidos nesse caso. Para o modelo em consideração, Van de Ven e Van Praag (1981) propõem incluir um termo de correção no modelo *probit* de interesse. Nesse caso, o procedimento de estimação sugerido por eles é o seguinte:

a) estimar um modelo *probit* para a inscrição no vestibular e obter estimativas de λ utilizando $\hat{\gamma}$, tal como no procedimento de Heckman descrito anteriormente;

b) estimar ρ através de um modelo de probabilidade linear para *S*, com $\hat{\lambda}$ entre os regressores, isto é, aplicar mínimos quadrados à $S_i = x_i'\beta + \rho\lambda_i + v_i$;

c) substituindo $\hat{\lambda}$ e $\hat{\rho}$ na equação (8), obter uma estimativa de τ^2; e

d) estimar os parâmetros β e ρ através de um *probit*, em que o índice é dado por $(x'\beta + \rho\hat{\lambda})/\hat{\tau} + \xi$. Note-se que a normalização por τ remove a heteroscedasticidade, de modo que $E(\xi^2 \mid I \geq 0) = 1$.

Embora esse procedimento represente apenas uma aproximação das estimativas de máxima verossimilhança, o alto custo computacional envolvido na otimização de (12) faz com que ele seja uma alternativa interessante. Isso é especialmente relevante caso essa seja uma boa aproximação, como argumentam Van de Ven e Van Praag (1981). De fato, como veremos a seguir, os resultados obtidos confirmam esse argumento.

A desvantagem do procedimento em dois estágios proposto por Van de Ven e Van Praag (1981) é que a utilização de uma estimativa de λ leva a alguma perda de eficiência, já que isto introduz uma fonte adicional de variabilidade. É importante notar que Van de Ven e Van Praag não consideram a correção da matriz de variância e co-variância para refletir o uso de $\hat{\lambda}$ no lugar de λ. Heckman (1979) considera tanto a correção para heteroscedasticidade quanto essa correção para $\hat{\lambda}$.

3 CONSTRUINDO A AMOSTRA CENSURADA

A aplicação do modelo já discutido depende da disponibilidade de uma amostra censurada, isto é, uma amostra em que todas as características dos candidatos potenciais são observadas, exceto a habilidade de cada um, representada por y_i. A amostra disponível é truncada. Toda informação disponível refere-se apenas ao grupo de indivíduos que optou por se inscrever no vestibular da Fuvest.

A solução proposta para esse problema é utilizar a PNAD para extrair informações relevantes sobre a população de interesse e construir uma amostra censurada

que combine os dados da PNAD e da Fuvest. Tal procedimento é possível, pois essas duas bases de dados possuem diversos quesitos em comum nos seus questionários, permitindo o casamento (*matching*) de indivíduos entre elas, como descrito a seguir.

Os dados disponibilizados pela Fuvest referem-se aos candidatos inscritos no vestibular 2000. A amostra original continha 149.240 observações. Eliminando-se os "treineiros", restaram 130.190 observações, cada uma representando um único candidato.[3] Além das respostas ao questionário socioeconômico apresentado aos candidatos no ato da inscrição ao vestibular, os dados disponíveis incluem o estado onde se localiza a residência, a idade dos candidatos e uma variável mostrando se o candidato matriculou-se em algum curso.

Do total de 352.393 observações disponíveis na amostra da PNAD de 1999, é possível obter informação sobre a educação dos pais apenas para aqueles que se declararam filhos dentro de cada um dos domicílios pesquisados, o que reduz a amostra para 159.199 observações.[4] Desse total de 159 mil observações, apenas 20.165 tinham escolaridade que as habilitasse a ingressar no ensino superior e idade entre 15 e 72 anos.[5] Essa amostra representa todos os brasileiros em condições de ingressar no ensino superior, inclusive aqueles que se inscreveram de fato no vestibular da Fuvest (pelo menos com segundo grau completo). Desse modo, para combinar os dados da Fuvest e da PNAD é preciso descobrir quais indivíduos estão representados nas duas amostras e ajustar os coeficientes de expansão associados a cada observação.

O primeiro passo para implementar esse ajuste nos coeficientes de expansão foi agregar, em cada amostra, os indivíduos com características observáveis idênticas. Quer dizer, indivíduos idênticos foram colocados em uma única observação, à qual está associado um fator de expansão indicando o número de indivíduos ali representados. No caso das observações no banco de dados da Fuvest, esse fator de expansão é, simplesmente, o número de indivíduos contidos em cada observação. Para as observações na amostra PNAD, por outro lado, a definição desse fator de expansão é mais problemática. Originalmente, cada observação da PNAD já está associada a um fator de expansão, de modo que, ao agregar essas observações, é preciso "agregar" também esses fatores de expansão. A alternativa utilizada

3. Treineiros são candidatos que, por ainda não terem concluído o segundo grau, se inscrevem em carreiras fictícias, apenas para testar seus conhecimentos.

4. O fato de utilizarmos apenas o universo de "filhos" amostrados pela PNAD significa que a amostra continua, de certa forma, truncada. Acreditamos, no entanto, que o viés decorrente desse truncamento não é significativo, já que a maior parte dos candidatos parece pertencer ao grupo "filhos".

5. Após excluir os "treineiros", essas são as idades mínima e máxima dos inscritos no vestibular 2000 da Fuvest. Decidimos utilizar esses limites ao extrair dados da PNAD para garantir uma certa compatibilidade entre os bancos. De qualquer forma, cabe destacar que o número de observações com menos de 16 anos (idade mínima com que ocorrem sucessos) é muito pequeno e praticamente não tem efeito sobre os resultados.

para essa agregação dos fatores de expansão foi a soma aritmética dos fatores individuais. Embora essa alternativa seja questionável do ponto de vista de representatividade estatística, ela nos parece suficientemente razoável como um indicador do peso relativo atribuído a cada observação no processo de estimação.

Para possibilitar o casamento (*matching*) de indivíduos presentes em cada amostra, as características utilizadas como critério devem estar disponíveis nos dois bancos de dados, quais sejam: idade, estado de origem, sexo, raça, educação da mãe e se o domicílio em questão possui máquina de lavar. De acordo com essas características, o banco da PNAD possuía 9.123 observações diferentes, enquanto a Fuvest possuía 4.327.

Finalmente, é preciso lembrar que as 9 mil observações extraídas da PNAD representam todos os candidatos potenciais, inclusive aqueles que são candidatos de fato, representados pelas 4 mil observações oriundas da Fuvest. Desse modo, para evitar um tipo de dupla contagem que faz com que as observações da PNAD tenham um peso relativo maior do que o verdadeiro, é preciso reduzir o fator de expansão associado aos indivíduos representados nas duas amostras. Devido à dificuldade em garantir a representatividade estatística usualmente associada aos fatores de expansão, optamos por simplesmente subtrair dos fatores de cada observação da PNAD o número de indivíduos com as mesmas características na Fuvest. Como destacado anteriormente, é importante lembrar que esses fatores de expansão "agregados" servem apenas como um indicador (ainda que imperfeito) do peso relativo de cada observação, em contraste com a idéia de representatividade estatística usualmente associada aos fatores de expansão.[6]

Ao final desse procedimento, passamos a ter um único banco de dados contendo informações sobre educação da mãe, idade, raça, sexo, estado de origem e uma *proxy* para a renda (item de conforto — máquina de lavar), além de uma variável binária indicando se o indivíduo se inscreveu ou não no vestibular para toda a população em condições de seguir estudos de nível superior. Além disso, para aqueles inscritos no vestibular, ainda estão disponíveis as informações levantadas no questionário socioeconômico e a variável binária que indica se o candidato obteve sucesso no vestibular.

4 DESCRIÇÃO DOS DADOS UTILIZADOS

O questionário socioeconômico apresentado aos candidatos no ato da inscrição ao vestibular é composto por 32 questões. Além do questionário, estão disponíveis

6. Outro fator a reforçar a visão de que os fatores de expansão corrigidos representam apenas o peso relativo das observações é que não é possível garantir a exatidão, mesmo dos fatores originais. Como destacado por um parecerista anônimo, o cálculo dos fatores de expansão da PNAD de 1999 é feito sobre projeções a partir de dados censitários e somente os resultados do Censo Demográfico 2000 podem garantir o grau de exatidão desses números.

informações sobre o estado onde se localiza a residência e a idade dos candidatos e uma variável indicando se o candidato matriculou-se na USP.

As Tabelas 1 e 2 apresentam um sumário das respostas para a maioria das perguntas do questionário socioeconômico — não foram incluídas apenas as seguintes perguntas: situação profissional do pai, situação profissional da mãe, como o candidato pretende se manter durante o curso, o número de TVs em cores, o número de equipamentos de som e a origem profissional do candidato. Além disso, algumas variáveis possuíam pouca variação, o que acarretaria problemas numéricos durante as estimações, e portanto foram agregadas ou eliminadas. Outras variáveis foram agregadas de forma que o banco de dados utilizado na estimação dos modelos contivesse um número razoável de variáveis. É importante notar que procuramos evitar a agregação de respostas nos casos em que os resultados mostraram-se sensíveis ao critério de agregação.

Os dados referentes à amostra Fuvest nas Tabelas 1 e 2 estão sintetizados em termos de percentagens do total de matriculados e do total de candidatos inscritos no vestibular. Por exemplo, temos que 56,1% dos matriculados cursaram o ensino fundamental apenas em escolas particulares, enquanto entre os inscritos essa percentagem é de apenas 43,0%. Os dados referentes à amostra PNAD estão resumidos na Tabela 2 como percentagens do total de "filhos" na amostra PNAD.

Analisando essas tabelas, percebemos que há indícios de uma diferenciação bem definida entre os grupos de pessoas inscritas e matriculadas, sugerindo a importância de algumas variáveis como determinantes do sucesso no vestibular. Em geral, características que aparecem com maior freqüência entre os matriculados do que entre os inscritos devem contribuir para o sucesso no vestibular, e vice-versa. Portanto, a interpretação dos dados apresentados é, de certa forma, direta. No entanto, vale a pena destacar alguns casos interessantes.

No caso das variáveis de educação do pai e educação da mãe, é possível ver que o grupo cujos pais têm educação superior, mestrado ou doutorado tem sua representação consideravelmente maior no grupo dos matriculados do que o observado entre os inscritos. Enquanto todos os outros grupos educacionais têm sua representação reduzida entre os matriculados, o grupo com educação superior, mestrado ou doutorado aumenta sua representação de 41,0% dos inscritos para 55,3% dos matriculados, no caso do pai, e de 34,0% para 47,3%, no caso da mãe.

Quanto ao sexo dos candidatos, também há uma inversão marcante, pois enquanto as mulheres são maioria nos inscritos (53,3%), elas passam a ser minoria no grupo daqueles que obtêm sucesso no vestibular (42,2%). Como último exemplo, citamos o período no qual os candidatos estudaram durante o ensino médio. Nesse caso, dois grupos têm sua representação aumentada, enquanto os demais diminuem sua participação. O grupo daqueles que estudaram em período

TABELA 1
SUMÁRIO DOS DADOS DA FUVEST
[em %]

	Matriculados	Inscritos		Matriculados	Inscritos
Participação em vestibulares			Pessoas sustentadas		
Fuvest e públicas	0,358	0,270	2 pessoas	0,057	0,056
Fuvest e outras	0,286	0,360	3 ou 4 pessoas	0,516	0,510
Já prestou	0,654	0,350	5 pessoas ou mais	0,413	0,410
Foi treineiro	0,172	0,130			
			Pessoas com renda		
Ensino fundamental			1 a 3 com o candidato	0,131	0,150
Só particular	0,561	0,430	1 a 3 sem o candidato	0,807	0,780
Maior parte pública	0,070	0,079	4 ou mais	0,033	0,041
Maior parte particular	0,079	0,074			
Metade em cada	0,033	0,038	Itens de conforto		
Exterior	0,003	0,002	1 empregado	0,361	0,300
			2 empregados ou mais	0,060	0,063
Ensino médio			Carro	0,884	0,840
Não comum	0,175	0,210	Videocassete	0,912	0,880
Federal	0,029	0,016	1 computador	0,637	0,580
Só particular	0,683	0,550	2 ou mais computadores	0,136	0,084
Maior parte pública	0,023	0,032	Máquina de lavar	0,949	0,933
Maior parte particular	0,033	0,038	Internet de casa	0,511	0,380
Metade em cada	0,007	0,012	Internet do trabalho	0,044	0,053
Exterior	0,004	0,002	Internet outros	0,161	0,180
Período diurno	0,727	0,650			
Noturno	0,069	0,150	Característica do candidato		
Maior parte diurno	0,078	0,088	Masculino	0,470	0,578
Maior parte noturno	0,028	0,045	Casado	0,036	0,035
			Outro estado civil	0,120	0,011

(continua)

(continuação)

	Matriculados	Inscritos		Matriculados	Inscritos
Curso preparatório			Superior completo	0,149	0,099
Menos de 6 meses	0,153	0,250	Superior incompleto	0,074	0,030
6 meses a 1 ano	0,240	0,240	Exerce atividade remunerada	0,260	0,310
1 ano a 1,5 ano	0,079	0,046	Não mora com a família	0,364	0,390
Mais de 1,5 ano	0,150	0,073	Negra	0,013	0,023
			Amarela	0,121	0,077
Educação dos pais			Parda	0,058	0,088
Pai			Indígena	0,005	0,006
Primário	0,096	0,150			
Ginásio	0,072	0,110	Região		
Colegial	0,229	0,250	Sudeste (exceto SP)	0,022	0,025
Superior ou mais	0,553	0,410	Sul	0,005	0,005
Mãe			Norte	0,001	0,000
Primário	0,110	0,160	Nordeste	0,003	0,001
Ginásio	0,099	0,140	Centro-Oeste	0,004	0,004
Colegial	0,273	0,300	Distrito Federal	0,002	0,002
Superior ou mais	0,473	0,340			

integral (7,1% - 9,8%) ou que realizaram seus estudos em período diurno (64,8% - 72,7%) aumenta, enquanto os demais grupos, principalmente dos que realizaram o ensino médio totalmente no período noturno (14,8% - 6,9%), diminuem sua participação.

A Tabela 3 apresenta o número de inscritos e de matriculados em cada uma das carreiras da Fuvest, assim como as relações candidato-vaga reunidas pela amostra. Por conveniência, e para facilitar a análise dos resultados, agregamos algumas carreiras de acordo com certas características comuns. Segundo a amostra disponível, a carreira mais concorrida é a de Jornalismo e Publicidade, com 53 candidatos por vaga, seguida por Polícia Militar (33,2) e Computação — São Carlos (31,6). As carreiras menos concorridas são as de Matemática, Física e Química, com sete candidatos por vaga, seguidas de Ciências Sociais, Filosofia, Geografia e História, agregadas como Humanas 01, com 7,5 candidatos por vaga. O curso de

TABELA 2
SUMÁRIO DOS DADOS DA PNAD *VERSUS* FUVEST
[em %]

	Fuvest		PNAD
	Matriculados	Inscritos	
Educação da mãe			
Primário	11,0	16,0	28,2
Ginásio	9,9	14,0	15,4
Colegial	27,3	30,0	19,6
Superior	47,3	34,0	10,2
Raça			
Branca	80,4	80,6	52,8
Negra	1,3	2,3	7,3
Amarela	12,1	7,7	1,5
Parda	5,8	8,8	38,2
Indígena	0,5	0,6	0,2
Sexo			
Masculino	47,0	57,8	47,8
Feminino	53,0	42,2	52,2
Idade			
15	0,0	0,1	0,2
16	2,1	3,5	1,4
17	23,4	33,2	6,2
18	28,7	25,1	9,4
19	16,6	13,1	10,2
20	7,7	6,9	9,7
20 a 25	12,5	10,9	33,4
25 a 30	4,4	3,5	14,3
30 a 35	2,4	1,8	7,2
Mais de 35	2,2	1,9	7,9

Fontes: Fuvest 2000 e PNAD de 1999.

TABELA 3
INSCRITOS *VERSUS* MATRICULADOS POR CARREIRA

Carreiras	Inscritos		Matriculados		Candidato-Vaga (a/b)
	(a)	(%)	(b)	(%)	
Medicina	14.342	11,0	482	6,2	29,8
Biologia	4.101	3,2	185	2,4	22,2
Odontologia	4.339	3,3	263	3,4	16,5
Enfermagem	3.389	2,6	239	3,1	14,2
Farmácia	3.381	2,6	185	2,4	18,3
Zootecnia e Veterinária	3.104	2,4	120	1,6	25,9
Psicologia	3.315	2,5	110	1,4	30,1
Fono, Fisio e Terapia Ocupacional	3.978	3,1	134	1,7	29,7
Direito	13.605	10,5	459	5,9	29,6
Economia, Administração e Contábil	10.780	8,3	670	8,7	16,1
Humanas 01	6.067	4,7	809	10,5	7,5
Humanas 02	8.597	6,6	1.073	13,9	8,0
Humanas 03	1.528	1,2	50	0,6	30,6
Artes	1.481	1,1	84	1,1	17,6
Arquitetura e Urbanismo	3.301	2,5	180	2,3	18,3
Jornalismo e Publicidade	9.536	7,3	180	2,3	53,0
Matemática, Física e Química	5.271	4,0	754	9,8	7,0
Computação — São Carlos	2.493	1,9	79	1,0	31,6
Escola Politécnica	10.729	8,2	857	11,1	12,5
Engenharia, outras	5.180	4,0	426	5,5	12,2
Polícia Militar	7.014	5,4	211	2,7	33,2
Outras	4.659	3,6	180	2,3	25,9
Total	130.190		7.730		16,8

Fonte: Fuvest 2000.

Humanas 01: Ciências Sociais, Filosofia, Geografia e História.

Humanas 02: Letras, Pedagogia, Biblioteconomia e Ciências da Terra.

Humanas 03: Curso de Audiovisual e Editoração.

Outras: Economia Agrícola, Educação Física, Esportes, Nutrição e Técnica Oftalmológica.

Medicina, curiosamente, aparece apenas em sexto lugar entre os mais concorridos, com praticamente a mesma relação candidato-vaga que Fonoaudiologia, Fisiologia e Terapia Ocupacional e Direito.

5 RESULTADOS OBTIDOS

O modelo apresentado na Seção 2 foi estimado utilizando as variáveis explicativas apresentadas na seção anterior.[7] Especificamente, o modelo para inscrição foi estimado utilizando a variável Idade e os seguintes conjuntos de variáveis *dummies*: Educação da Mãe — referência analfabeto ou primário incompleto; Raça — referência branca; Sexo — referência feminino; Máquina de Lavar — referência não possui; e região de origem do candidato — referência Estado de São Paulo.

Para o modelo de sucesso no vestibular, foram utilizadas todas as variáveis anteriores, exceto a região de origem dos candidatos. Além dessas variáveis, foram utilizados os seguintes conjuntos de variáveis *dummies*: Participação em Vestibulares — referência inscrito apenas Fuvest; Ensino Fundamental — referência cursou apenas escola pública; Ensino Médio — referência cursou o colegial comum, em escola pública no período noturno; Curso Preparatório — referência não cursou; Pessoas Sustentadas — referência apenas uma; Pessoas com Renda — referência candidato é o único responsável pela renda familiar;[8] Pessoas Sustentadas — referência apenas o candidato; Itens de Conforto — referência nenhum empregado, não tem carro, videocassete, computador ou máquina de lavar e não tem acesso à internet; Carreiras — referência Medicina; e Características do Candidato — referência mulher, branca, solteira, apenas colegial completo, sem atividade remunerada e que irá morar com a família durante o curso.

A Tabela 4 apresenta os resultados obtidos através de cada um dos métodos de estimação discutidos. As primeiras duas colunas apresentam os coeficientes estimados através do modelo *probit* com correção e os respectivos desvios-padrão.[9] As três colunas seguintes referem-se aos resultados obtidos através do método de máxima verossimilhança: coeficientes estimados, desvios-padrão e efeitos marginais. Nas últimas três colunas são apresentados os resultados para o modelo *probit* sem correção: coeficientes estimados, desvios-padrão e efeitos marginais. Para facilitar a análise dos resultados obtidos, as variáveis explicativas foram agrupadas por tipo, como indicado pelos títulos.

---

7. É importante notar que existem variáveis potencialmente importantes que não estão disponíveis e, portanto, não foram incluídas no modelo. Isso pode, é claro, levar aos problemas clássicos de omissão de variáveis relevantes.

8. Cabe lembrar que não há na amostra PNAD indivíduos classificados como os únicos responsáveis pela renda familiar. Como discutido anteriormente, isso é conseqüência da inclusão apenas daqueles que se declararam "filhos" na PNAD (ver nota de rodapé 4).

9. Como os efeitos marginais para o modelo *probit* com correção são praticamente iguais àqueles obtidos para as estimativas de máxima verossimilhança, apresentamos apenas os resultados para esse último caso.

TABELA 4
COEFICIENTES E EFEITOS MARGINAIS ESTIMADOS

Variáveis	*Probit* com correção		Máxima verossimilhança			*Probit* simples		
	Coeficiente	Desvio-padrão	Coeficiente	Desvio-padrão	Efeito marginal	Coeficiente	Desvio-padrão	Efeito marginal
Participação em vestibulares								
Fuvest e públicas	0,181***	0,017	0,178***	0,02	0,013	0,166***	0,017	0.018
Fuvest e outras	–0,109***	0,018	–0,109***	0,02	–0,006	–0,123***	0,018	–0.011
Já prestou	0,540***	0,017	0,533***	0,02	0,037	0,529***	0,017	0.056
Foi treineiro	0,431***	0,020	0,426***	0,02	0,035	0,421***	0,020	0.051
Ensino fundamental								
Só particular	0,122***	0,021	0,121***	0,02	0,008	0,124***	0,021	0.012
Maior parte pública	0,030	0,029	0,030	0,03	0,002	0,030***	0,029	0.003
Maior parte particular	0,048*	0,029	0,048*	0,03	0,003	0,052*	0,029	0.005
Metade em cada	0,053	0,038	0,052	0,04	0,003	0,055	0,039	0.005
Exterior	0,062	0,153	0,061	0,15	0,004	0,058	0,153	0.006
Ensino médio								
Não comum	–0,068***	0,020	–0,068***	0,02	–0,004	–0,073***	0,020	–0.007
Federal	0,471***	0,047	0,466***	0,05	0,036	0,483***	0,047	0.056

(continua)

(continuação)

Variáveis	*Probit* com correção		Máxima verossimilhança			*Probit* simples		
	Coeficiente	Desvio-padrão	Coeficiente	Desvio-padrão	Efeito marginal	Coeficiente	Desvio-padrão	Efeito marginal
Só particular	0,204***	0,021	0,202***	0,02	0,013	0,207***	0,021	0.020
Maior parte pública	–0,016	0,044	–0,015	0,04	–0,001	–0,011	0,044	–0.001
Maior parte particular	0,068*	0,039	0,067*	0,04	0,004	0,072*	0,039	0.006
Metade em cada	–0,114	0,078	–0,113	0,08	–0,005	–0,113	0,078	–0.009
Exterior	0,443***	0,132	0,437***	0,13	0,033	0,442***	0,132	0.050
Período diurno	–0,112***	0,024	–0,111***	0,02	–0,008	–0,116***	0,024	–0.013
Noturno	–0,357***	0,034	–0,353***	0,03	–0,022	–0,362***	0,033	–0.035
Maior parte diurno	–0,174***	0,033	–0,172***	0,03	–0,012	–0,178***	0,032	–0.019
Maior parte noturno	–0,253***	0,044	–0,251***	0,04	–0,017	–0,258***	0,044	–0.027
Curso preparatório								
Menos que 6 meses	–0,025	0,020	–0,024	0,02	–0,001	–0,023	0,020	–0.002
6 meses a 1 ano	0,302***	0,019	0,298***	0,02	0,020	0,303***	0,019	0.031
1 ano a 1,5 ano	0,360***	0,030	0,357***	0,03	0,025	0,363***	0,030	0.038
Mais de 1,5 ano	0,562***	0,026	0,556***	0,03	0,045	0,563***	0,026	0.067

(continua)

(continuação)

Variáveis	*Probit* com correção		Máxima verossimilhança			*Probit* simples		
	Coeficiente	Desvio-padrão	Coeficiente	Desvio-padrão	Efeito marginal	Coeficiente	Desvio-padrão	Efeito marginal
Educação dos pais								
Pai								
Primário	–0,012	0,038	–0,012	0,04	–0,001	–0,012	0,038	–0.001
Ginásio	–0,020	0,041	–0,020	0,04	–0,001	–0,020	0,041	–0.002
Colegial	0,032	0,038	0,031	0,04	0,002	0,032	0,038	0.003
Superior ou mais	0,105***	0,039	0,104***	0,04	0,007	0,105***	0,039	0.011
Mãe								
Primário	–0,015	0,039	–0,013	0,04	–0,001	–0,003	0,039	0.000
Ginásio	–0,016	0,042	–0,017	0,04	–0,001	–0,044	0,042	–0.004
Colegial	0,043	0,041	0,040	0,04	0,002	–0,009	0,040	–0.001
Superior ou mais	0,260***	0,045	0,252***	0,04	0,017	0,145***	0,041	0.015
Pessoas com renda								
1 a 3 com o candidato	–0,101**	0,051	–0,100**	0,05	–0,007	–0,099*	0,051	–0.011
1 a 3 sem o candidato	–0,108**	0,054	–0,107**	0,05	–0,008	–0,105**	0,054	–0.011
4 ou mais	–0,116*	0,062	–0,114*	0,06	–0,008	–0,114***	0,062	–0.012

(continua)

(continuação)

Variáveis	*Probit* com correção		Máxima verossimilhança			*Probit* simples		
	Coeficiente	Desvio-padrão	Coeficiente	Desvio-padrão	Efeito marginal	Coeficiente	Desvio-padrão	Efeito marginal
Pessoas sustentadas								
2 pessoas	0,153**	0,069	0,152**	0,07	0,009	0,159**	0,069	0.014
3 ou 4 pessoas	0,159**	0,064	0,157**	0,06	0,009	0,165***	0,065	0.015
5 pessoas ou mais	0,154**	0,065	0,152**	0,06	0,009	0,161***	0,065	0.015
Itens de conforto								
Um empregado	–0,028*	0,016	–0,027*	0,02	–0,002	–0,026*	0,016	–0.003
2 empregados ou mais	–0,096***	0,029	–0,094***	0,03	–0,006	–0,092***	0,029	–0.009
Carro	–0,028	0,022	–0,027	0,02	–0,002	–0,029	0,022	–0.003
Videocassete	0,012	0,024	0,012	0,02	0,001	0,011	0,024	0.001
1 computador	0,043**	0,019	0,042**	0,02	0,003	0,041**	0,019	0.004
2 ou mais computadores	0,172***	0,028	0,170***	0,03	0,012	0,169***	0,027	0.018
Máquina de lavar	0,064*	0,034	0,060*	0,03	0,004	–0,024	0,032	–0.002
Internet de casa	0,095***	0,018	0,094***	0,02	0,006	0,096***	0,018	
Internet do trabalho	0,014	0,036	0,013	0,04	0,001	0,013	0,036	0.001
Internet outros	0,030	0,020	0,030*	0,02	0,002	0,031***	0,020	0.003

(continua)

(continuação)

Variáveis	*Probit* com correção		Máxima verossimilhança			*Probit* simples		
	Coeficiente	Desvio-padrão	Coeficiente	Desvio-padrão	Efeito marginal	Coeficiente	Desvio-padrão	Efeito marginal
Carreiras								
Biologia	0,356***	0,046	0,352***	0,05	0,012	0,351***	0,046	0.019
Odontologia	0,496***	0,042	0,490***	0,04	0,019	0,491***	0,042	0.030
Enfermagem	0,937***	0,045	0,926***	0,05	0,054	0,929***	0,045	0.083
Farmácia	0,518***	0,047	0,512***	0,05	0,020	0,512***	0,047	0.032
Zootecnia e Veterinária	0,355***	0,053	0,350***	0,05	0,012	0,647***	0,032	0.019
Psicologia	0,405***	0,055	0,400***	0,05	0,014	0,587***	0,044	0.023
Fono, Fisio e Terapia Ocupacional	0,416***	0,050	0,411***	0,05	0,015	0,408***	0,050	0.023
Direito	0,334***	0,034	0,330***	0,03	0,011	0,329***	0,034	0.018
Economia, Administração e Contábil	0,666***	0,033	0,658***	0,03	0,030	0,660***	0,033	0.047
Humanas 01	1,214***	0,035	1,200***	0,04	0,089	1,206***	0,035	0.131
Humanas 02	1,422***	0,034	1,405***	0,03	0,123	1,414***	0,034	0.177
Humanas 03	0,333***	0,077	0,329***	0,08	0,011	0,333***	0,077	0.018
Artes	0,789***	0,064	0,780***	0,06	0,039	0,783***	0,064	0.062
Arquitetura e Urbanismo	0,531***	0,047	0,525***	0,05	0,021	0,525***	0,047	0.033

(continua)

(continuação)

Variáveis	*Probit* com correção		Máxima verossimilhança			*Probit* simples		
	Coeficiente	Desvio-padrão	Coeficiente	Desvio-padrão	Efeito marginal	Coeficiente	Desvio-padrão	Efeito marginal
Jornalismo e Propaganda/Publicidade	0,210***	0,043	0,207***	0,042	0,006	0,205***	0,043	0.010
Matemática, Física e Química	1,289***	0,035	1,274***	0,04	0,100	1,281***	0,035	0.147
Computação — São Carlos	0,276***	0,062	0,273***	0,06	0,008	0,268***	0,062	0.014
Escola Politécnica	0,649***	0,032	0,642***	0,03	0,028	0,483***	0,047	0.046
Engenharia, outras	0,757***	0,038	0,748***	0,04	0,037	0,751***	0,038	0.058
Polícia Militar	0,606***	0,044	0,598***	0,04	0,025	0,587***	0,044	0.039
Outras	0,491***	0,047	0,486***	0,05	0,018	0,483***	0,047	0.030
Característica do candidato								
Homem	0,263***	0,015	0,260	0,02	0,017	0,267	0,015	0.027
Casado	0,000	0,046	0,000	0,05	0,000	0,003	0,079	0.000
Outro estado civil	0,060	0,069	0,060	0,07	0,004	0,062	0,069	0.007
Superior incompleto	0,065***	0,023	0,064***	0,02	0,004	0,064***	0,023	0.006
Superior completo	0,536***	0,040	0,529***	0,04	0,050	0,532***	0,040	0.072
Exerce atividade remunerada	–0,124***	0,022	–0,122***	0,02	–0,008	–0,126***	0,022	–0.012
Não irá morar com a família	–0,115***	0,015	–0,112***	0,02	–0,007	0,095***	0,015	–0.009

(continua)

(continuação)

Variáveis	*Probit* com correção		Máxima verossimilhança			*Probit* simples		
	Coeficiente	Desvio-padrão	Coeficiente	Desvio-padrão	Efeito marginal	Coeficiente	Desvio-padrão	Efeito marginal
Negra	–0,148***	0,056	–0,151***	0,06	–0,009	–0,186***	0,056	–0.017
Amarela	0,206***	0,025	0,198***	0,02	0,015	0,129***	0,021	0.014
Parda	–0,080***	0,028	–0,079***	0,03	–0,005	–0,062***	0,028	–0.006
Indígena	0,126	0,107	0,108	0,10	0,008	–0,152	0,099	–0.014
Idade	–0,048***	0,012	–0,046***	0,01		–0,009	0,010	
Idade ao quadrado	0,001***	0,000	0,001***	0,00		0,000	0,000	
Constante	–2,835***	0,170	–2,818***	0,17		–2,920***	0,169	
Razão de Mills	0,185***	0,027	0,176***	0,02				

*** Nível de significância menor que 1%.

** Significativo ao nível de 5%.

* Significativo ao nível de 10%.

Os efeitos marginais apresentados na Tabela 4 foram calculados seguindo o procedimento proposto por Wooldridge (2001, p. 467). Para as variáveis binárias o procedimento para estimar o efeito de uma variável x_k foi o seguinte: *a*) calculamos a probabilidade de inscrição e sucesso no vestibular para cada ponto da amostra quando $x_{ik} = 1$ para todo $i = 1, ..., N$; *b*) da mesma forma, calculamos aquela probabilidade quando $x_{ik} = 0$ para todas as observações; e *c*) estimamos o efeito marginal como a diferença entre as médias dessas probabilidades na amostra. Para a variável Idade (a única variável contínua), o procedimento adotado foi semelhante, mas com x_k variando de 15 a 72 em inteiros.

De maneira geral, os números apresentados na Tabela 4 mostram que, mesmo sendo uma aproximação, o modelo *probit* com correções para viés de seleção e heteroscedasticidade leva a coeficientes muito próximos daqueles obtidos por máxima verossimilhança. Tanto a significância dos parâmetros quanto os valores encontrados para cada um dos coeficientes permanecem praticamente inalterados.

Como discutido anteriormente, o viés de seleção existe sempre que $\rho \neq 0$. Como esse parâmetro é estimado tanto pelo *probit* com correção quanto pela máxima verossimilhança, podemos testar a presença do viés através do teste de significância desses coeficientes. No caso do *probit* com termo de correção, a estimativa r é dada pelo coeficiente estimado para a Razão de Mills Invertida. No caso da estimação por máxima verossimilhança, r é estimado diretamente e não como o coeficiente de um regressor. No entanto, para facilitar a comparação e a apresentação, a estimativa de máxima verossimilhança foi incluída também na linha correspondente à Razão de Mills Invertida. Os resultados obtidos mostram que esse coeficiente é positivo e significativamente diferente de 0, revelando que a hipótese nula de independência das equações é rejeitada em ambos os modelos.

Começando com as variáveis relacionadas à participação do candidato em vestibulares, constatamos que os candidatos inscritos também em vestibulares de escolas particulares têm menos chances de obter sucesso no vestibular (–0,6%) com relação àqueles que estão inscritos apenas na Fuvest. Os candidatos que prestam vestibular em outras instituições públicas além da Fuvest, por outro lado, têm probabilidade de sucesso maior (1,3%). Uma possível interpretação para esses resultados é que essas variáveis capturam a expectativa do candidato com relação a suas chances no vestibular. Assim, o indivíduo que não está tão confiante em seu desempenho nos vestibulares de escolas públicas inscreve-se nos de escolas particulares.

A participação em vestibulares no passado tem um efeito positivo sobre a probabilidade de sucesso. A participação como treineiro no ano anterior aumenta a probabilidade de sucesso em cerca de 3,5%, ao passo que a participação como candidato tem um efeito de 3,7%. Esse resultado, como era esperado, mostra que experiência e persistência têm um efeito positivo sobre a probabilidade de sucesso.

Os próximos dois grupos de variáveis referem-se ao tipo de escola onde o candidato realizou seus estudos no ensino fundamental e no ensino médio. Em ambos os casos, o fato de ter estudado tanto em escolas públicas quanto particulares não proporciona nenhum efeito significativo sobre aqueles que só estudaram em escolas públicas. No caso do ensino fundamental, apenas aqueles que estudaram só em escolas particulares têm maiores chances de sucesso (0,8%). No caso do ensino médio, por outro lado, além daqueles que só estudaram em escolas particulares (1,3%), têm alguma vantagem os que estudaram em escolas federais (3,6%) ou no exterior (3,3%).

Ainda com relação ao ensino médio, além do efeito público *versus* privado, é possível avaliar também o efeito da especialização recebida (comum, magistério, técnico etc.) e o período. Os resultados obtidos mostram que a especialização durante o ensino médio tem um impacto negativo na probabilidade de sucesso (–0,4%). Quanto ao período, constatamos que estudar em período integral durante o ensino médio é a melhor opção, seguido pelo período diurno durante toda a duração do ensino médio ou apenas em parte (–0,8% e –1,2%, respectivamente). Estudar no período noturno durante todo o ensino médio ou em parte dele tem um impacto negativo maior do que quaisquer outras opções disponíveis (–2,2% e –1,7%, respectivamente). Esse resultado revela que, mesmo controlando pelas demais características, o ensino noturno reduz a probabilidade de sucesso do candidato no vestibular.

A seguir, temos as variáveis que identificam o efeito do número de semestres que o candidato se preparou em "cursos pré-vestibulares". Vemos que aqueles que se preparam por mais de um semestre e menos de um ano e aqueles que se prepararam entre dois e três semestres têm maiores chances de sucesso quando comparados àqueles que se prepararam durante menos de um semestre (2,0% e 2,5%, respectivamente), ou que nunca freqüentaram cursos preparatórios. O maior impacto na probabilidade de sucesso é para aqueles que freqüentaram cursos preparatórios por mais de três semestres (4,5%).

O próximo grupo de variáveis refere-se à educação dos pais. Os candidatos cujos pais possuem nível superior completo, mestrado ou doutorado têm probabilidade de sucesso maior do que aqueles cujos pais não têm instrução alguma ou apenas primário incompleto. Um aspecto que chama a atenção é que o efeito de a mãe possuir nível superior, mestrado ou doutorado é maior do que o efeito nível de escolaridade do pai (1,7% contra 0,7%). Vale ressaltar que, uma vez que não dispomos da renda familiar, a educação dos pais pode estar captando parte dos efeitos da variável renda.

Os próximos três grupos referem-se diretamente às condições de renda familiar dos candidatos. Começando com o número de pessoas que contribuem

para a renda familiar, verificamos que os candidatos que são os únicos responsáveis pela renda familiar (grupo de referência para esse conjunto de variáveis) têm probabilidade maior de sucesso. Esse resultado é, de certa forma, inesperado. Quando há outras pessoas responsáveis pela renda familiar, o efeito marginal é praticamente o mesmo, independentemente da condição do candidato (0,7% quando ele contribui e 0,8% quando não). Finalmente, é interessante notar que esse conjunto de variáveis é o único que passa de não-significativo para significativo ao corrigirmos o viés de seleção.

Com relação ao número de pessoas sustentadas pela renda familiar, a existência de mais de uma pessoa dependente da renda familiar contribui positivamente para o sucesso do candidato, cerca de 0,9%. Em outras palavras, os candidatos que moram sozinhos têm menos chances no vestibular.

Os itens de conforto utilizados como *proxies* para a renda familiar do candidato não se apresentaram significativos, com exceção das respostas referentes à presença de computador no domicílio e de mais de dois empregados domésticos. Interessantemente, a presença de mais de dois empregados mensalistas influencia, de modo negativo, a probabilidade de sucesso, quando comparados àqueles que não possuem empregado mensalista, em média –0,6%. A presença de computador na residência, por sua vez, tem um efeito de cerca de 0,3%, no caso de um computador, e de 1,2%, no caso de dois ou mais computadores. O acesso à internet se mostra significativo para aqueles que acessam o serviço de casa (0,6%), sendo não-significativo nos casos em que o acesso é feito de outros locais.

A presença de máquina de lavar no domicílio, apesar de não ser estatisticamente significativa, merece destaque. Ao corrigir o viés de seleção, o coeficiente desse item de conforto tem seu nível de significância reduzido consideravelmente e a direção do efeito se inverte, passando de negativo para positivo. Note-se que esse é o único item de conforto incluído ao modelarmos a inscrição, de modo que essa alteração pode ser vista como uma indicação de que o efeito obtido no *probit* simples incorpora parte do efeito de decidir pela inscrição.

O conjunto de variáveis referentes à carreira pela qual o candidato fez opção reflete tanto as diferentes relações candidato-vaga quanto as notas de corte associadas a cada uma das carreiras. Os resultados obtidos mostram que o candidato a uma vaga no curso de Medicina é o que tem a menor chance de sucesso. O curso de Medicina é seguido de Jornalismo e Publicidade e Propaganda (agregados). Muito embora Engenharia seja tradicionalmente tido como um curso com maior dificuldade de ingresso, é curioso notar que os candidatos a Escola Politécnica têm maior probabilidade de sucesso que aqueles que se candidatam a Farmácia ou Psicologia, por exemplo. Esse resultado, provavelmente, se relaciona com o fato de que sob a variável "Poli" estão todos os cursos ministrados na Escola

Politécnica, de modo que tanto os cursos de Engenharia, pouco procurados, como aqueles mais disputados estão representados pela mesma variável.

Um ponto importante a ser ressaltado é que o efeito marginal das variáveis de carreira foi bastante alterado após controlarmos o viés de seleção, pois em média o efeito marginal foi reduzido em 54%. Muito embora o coeficiente de correlação encontrado entre a equação que modela a inscrição e a que modela o resultado obtido no vestibular seja baixo, verificamos que a seleção gera um impacto considerável nos efeitos marginais das variáveis estudadas.

Finalmente, há o grupo de variáveis com características do indivíduo. Candidatos do sexo masculino têm chance de sucesso ligeiramente maior que candidatos do sexo feminino (1,7%), enquanto o estado civil dos candidatos não tem efeito significativo. O desafio do vestibular é mais facilmente superado por aqueles que já freqüentaram outro curso superior: 5,0% se concluíram e 0,4%, caso contrário. Como esperado, exercer atividade remunerada diminui as chances de obter sucesso no vestibular (cerca de –0,8%). Aqueles que esperam morar com a família, caso sejam aprovados, têm maiores chances de obter sucesso (0,7%).

Raça é uma variável significativa para todos os grupos, exceto o indígena (composto por poucas observações). Os resultados mostram que os brancos têm vantagem quando comparados aos negros (–0,9%) e pardos (–0,5%), mas estão em desvantagem quando comparados aos de etnia amarela (1,5%). Esse é um resultado inesperado, pois, controlando-se por todas as outras características, esperávamos que raça não influenciasse o resultado do exame, uma vez que o vestibular não possui nenhum componente de subjetividade, o que exclui discriminação pura. Esse resultado, provavelmente, é fruto de efeitos de renda não captados pelos indicadores de conforto presentes na regressão.

O efeito idade é ilustrado no gráfico a seguir e mostra que a probabilidade de sucesso declina com a idade do indivíduo até atingir o seu mínimo por volta dos 41 anos e, após isso, aumenta até os 72 anos. É importante notar que no *probit* simples tanto a variável idade quanto o quadrado da idade eram estatisticamente não-significativos, porém, ao se controlar o viés de seleção, essas variáveis passam a ser significativas. O declínio na probabilidade de sucesso está ligado, provavelmente, ao fato de que com a idade, em geral, as pessoas vão se distanciando dos estudos, e passar muito tempo longe dos estudos diminui a probabilidade de sucesso.

Para finalizar a análise dos resultados obtidos, resta examinar a capacidade de previsão do modelo estimado. A Tabela 5 apresenta a classificação das observações de acordo com os modelos estimados *vis-à-vis* a classificação observada. A classificação pelos modelos foi feita utilizando-se um "valor de corte" igual a 7%, que

EFEITO MARGINAL — IDADE

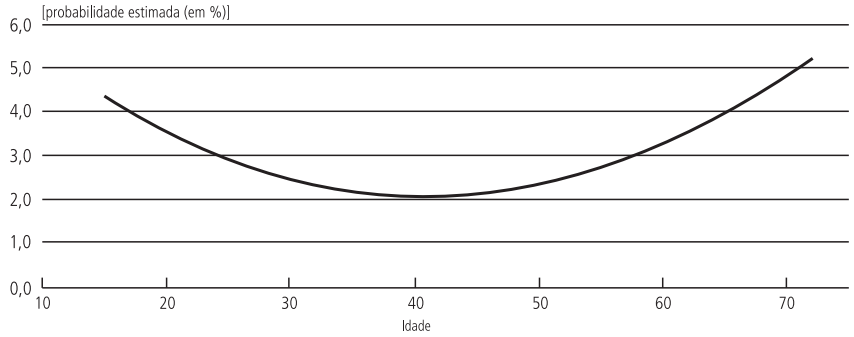

TABELA 5
PREVISÕES *VERSUS* OBSERVADO

	Com correção			Sem correção		
	S = 0	*S* = 1	Total	*S* = 0	*S* = 1	Total
	95.624	3.767	99.391	83.275	2.164	85.439
	11.897	3.201	15.098	24.246	4.804	29.050
Total	107.521	6.898	114.489	107.521	6.898	114.489
% correto	88,94	45,94	86,32	77,45	68,94	76,93
% incorreto	11,06	54,06	13,68	22,55	31,06	23,07

é ligeiramente superior à freqüência relativa de matriculados na amostra (6,48%). Em outras palavras, todos os candidatos para os quais a probabilidade de sucesso dada pelos modelos é maior do que a freqüência relativa de matriculados foram classificados como sucesso.

O *probit* simples, como podemos ver, prevê corretamente 68,9% dos sucessos e 77,5% dos insucessos. O modelo com correção para viés de seleção, por sua vez, prevê corretamente 45,9% dos sucessos e 88,9% dos insucessos. Em média, o modelo corrigido prevê os valores da variável dependente de maneira correta para 86,3% das observações, enquanto o *probit* simples prevê apenas 76,9% das observações. Logo, podemos concluir que o modelo com correção para viés de seleção apresenta um desempenho um pouco melhor que o modelo *probit* sem correção.

6 CONCLUSÃO

O objetivo deste trabalho era utilizar os dados da Fuvest para verificar quais são os determinantes do acesso à USP, levando em consideração o viés de seleção inerente à natureza dos dados. Foram utilizados métodos de estimação apropriados para testar a existência desse tipo de viés e corrigir seus efeitos. Através desse procedimento constatamos que a hipótese de ausência de um viés de seleção na amostra é rejeitada.

Em linhas gerais, os resultados obtidos são semelhantes aos encontrados em outros trabalhos similares, como Siano (1977), Freitas (1979) e Oliveira (1980), no que se refere a quais são os determinantes do sucesso no vestibular. Assim como aqueles autores, constatamos que os fatores mais importantes são a educação dos pais e o tipo de escolas que os candidatos freqüentaram. Além disso, constatamos uma marcante diferença de dificuldade entre as diversas carreiras.

No entanto, a correção do modelo levou a estimativas de efeitos menores do que aqueles estimados sem levar em conta o viés de seleção. Como o nível de significância de cada um desses parâmetros foi pouco afetado pela correção, podemos concluir que o efeito da correção do viés de seleção é essencialmente quantitativo.

Essa mudança quantitativa nos resultados é muito importante em um contexto em que se discute definição de políticas de acesso ao ensino superior gratuito. Em geral, o aspecto qualitativo dos resultados vai ao encontro das expectativas geradas pelo senso comum. Por exemplo, é esperado que candidatos oriundos de escolas públicas tenham mais dificuldades para obter sucesso no vestibular. Mas uma política de redução das notas de corte para esses candidatos depende de maneira fundamental da quantificação de sua "desvantagem" com relação a outros candidatos.

Analisando as variáveis com efeito positivo sobre a probabilidade de sucesso no vestibular, verificamos que os cursos pré-vestibulares e mães com nível superior têm efeitos de relativa magnitude. A importância dos cursos pré-vestibulares sugere que ações no sentido de oferecer acesso a eles estão entre as mais efetivas. O grau de escolaridade da mãe, por outro lado, mostra que há um efeito de longo prazo a ser considerado na avaliação de qualquer política de acesso ao ensino superior.

Quanto ao tipo de escola freqüentado, verificamos que a desvantagem público *versus* privado já se pronuncia durante o ensino fundamental, embora em menor escala que no ensino médio. A maior vantagem é daqueles que cursaram o ensino médio em escolas federais ou no exterior, sugerindo que, na média, a diferença qualitativa entre as escolas particulares e públicas é bem menor do que aquela entre as particulares e as federais ou estrangeiras.

Finalmente, é importante destacar que o efeito marginal da raça do candidato é relativamente pequeno, embora significativo. Uma vez que controlamos pelas diferentes carreiras e outras características socioeconômicas, esse efeito deve refletir alguma dimensão da renda não-capturada pelas *proxies* utilizadas ou, talvez, possíveis diferenças culturais não-observáveis. De qualquer forma, o efeito negativo atribuído pelo modelo às raças negra e parda com relação à branca (0,9% e 0,5%, respectivamente) quase não justifica a introdução de cotas raciais proporcionalmente elevadas.

ABSTRACT

The access to higher education is commonly pointed as a critical issue for the alleviation of the income inequality in Brazil. This article presents an analysis of the determinants of the access to state universities, which is potentially useful for analyzing public policies for increasing the access to higher education. Using data from vestibular 2000 by Fuvest and from PNAD 1999, an econometric model has been estimated with correction for the selection bias arising from the fact that only those individuals taking the vestibular exams are observed.

BIBLIOGRAFIA

AMUEDO-DORANTES, C., KIMMEL, J. *Do college educated women in the United States delay fertility as a means of reducing the motherhood wage penalty*? Western Michigan University, Department of Economics, 2003 (Working Paper, 07-03).

BARROS, R. P., MENDONÇA, R. Os determinantes da desigualdade no Brasil. *A Economia Brasileira em Perspectiva*, v. 2. Rio de Janeiro: IPEA, 1996.

BECKER, G. *Human capital*. New York: NBER, 1964.

FERNANDES, R., MENEZES-FILHO, N. A. A evolução da desigualdade no Brasil metropolitano entre 1983 e 1997. *Estudos Econômicos*, v. 30, n. 4, 2000.

FERNANDES, R., NARITA, R. D. T. Instrução superior e mercado de trabalho no Brasil. *Economia Aplicada*, v. 5, n. 1, p. 7-32, 2001.

FREEMAN, R. B. Demand for education. In: ASHENFELTER, O., LAYARD, R. (eds.). *Handbook of Labor Economics*, v. 1, p. 357-386. Amsterdan: North-Holland, 1986.

FREITAS, L. P. *Os fatores condicionantes e determinantes do êxito no vestibular — o caso da Fundação Universidade Federal do Piauí em 1978*. Rio de Janeiro: Pontifícia Universidade Católica, 1979 (Dissertação de Mestrado).

GHIDONI, M. Determinants of young Europeans' decision to leave the parental household. *Royal Economic Society Annual Conference,* n. 85, 2002.

GREENE, W. *Econometric analysis*. 4th ed. New Jersey: Prentice-Hall, 2000.

HECKMAN, J. Sample selection bias as a specification error. *Econometrica*, v. 47, n. 1, p. 153-161, 1979.

OLIVEIRA, J. H. *Determinantes do sucesso no vestibular 79 da Universidade Federal de Viçosa — Minas Gerais*. Rio de Janeiro: Pontifícia Universidade Católica, 1980 (Dissertação de Mestrado).

PAINTER, G. Tenure choice with sample selection: differences among alternative samples. *Journal of Housing Economics*, v. 9, n. 3, p. 197-213, 2000.

PAINTER, G., GABRIEL, S. A., MYERS, D. Race, immigrant status, and housing tenure choice. *Journal of Urban Economics*, v. 49, n. 1, p. 150-167, 2001.

SIANO, L. M. F. *Determinantes do êxito do vestibular na Universidade Federal do Espiríto Santo (UFES)*. Rio de Janeiro: Pontifícia Universidade Católica, 1977 (Dissertação de Mestrado).

SPENCE, M. Job market signaling. *Quaterly Journal of Economics*, v. 87, n. 3, p. 355-374, 1973.

VAN DE VEN, W. P. M. M., VAN PRAAG, B. M. S. The demand for deductibles in private health insurance: a probit model with sample selection. *Journal of Econometrics*, v. 17, n. 2, p. 229-252, 1981.

WOOLDRIDGE, J. M. *Econometric analysis of cross section and panel data*. Cambridge: MIT Press, 2001.

(Originais recebidos em maio de 2003. Revistos em abril de 2004.)