TEXTO PARA DISCUSSÃO Nº 1240

AVALIAÇÃO DE UMA *PROXY* PARA A IDADE DA FIRMA UTILIZANDO AMOSTRAGEM COMPLEXA

Gustavo Costa
Patrick Alves
Mirian Bittencourt
Kátia Araújo
Hélyo Doyle

Brasília, dezembro de 2006

TEXTO PARA DISCUSSÃO Nº 1240

AVALIAÇÃO DE UMA *PROXY* PARA A IDADE DA FIRMA UTILIZANDO AMOSTRAGEM COMPLEXA*

Gustavo Costa**
Patrick Alves**
Mirian Bittencourt**
Kátia Araújo**
Hélyo Doyle**

Brasília, dezembro de 2006

* Os autores agradecem as sugestões de Fernando Freitas, consultor estatístico da Diretoria de Estudos Setoriais (Diset) do Ipea.
** Consultores Estatísticos da Diretoria de Estudos Setoriais (Diset) do Ipea.

Governo Federal

Ministério do Planejamento, Orçamento e Gestão

Ministro – Paulo Bernardo Silva
Secretário-Executivo – João Bernardo de Azevedo Bringel

ipea **Instituto de Pesquisa Econômica Aplicada**

Fundação pública vinculada ao Ministério do Planejamento, Orçamento e Gestão, o Ipea fornece suporte técnico e institucional às ações governamentais – possibilitando a formulação de inúmeras políticas públicas e programas de desenvolvimento brasileiro – e disponibiliza, para a sociedade, pesquisas e estudos realizados por seus técnicos.

URL: http://www.ipea.gov.br

Ouvidoria: http://www.ipea.gov.br/ouvidoria

ISSN 1415-4765

JEL C4, C13, C14, C87, J10

TEXTO PARA DISCUSSÃO

Publicação cujo objetivo é divulgar resultados de estudos direta ou indiretamente desenvolvidos pelo Ipea, os quais, por sua relevância, levam informações para profissionais especializados e estabelecem um espaço para sugestões.

A produção editorial desta publicação contou com o apoio financeiro do Banco Interamericano de Desenvolvimento (BID), via Programa Rede de Pesquisa e Desenvolvimento de Políticas Públicas – Rede-Ipea, o qual é operacionalizado pelo Programa das Nações Unidas para o Desenvolvimento (Pnud), por meio do Projeto BRA/04/052.

SUMÁRIO

SINOPSE

A importância dada ao tempo de atuação de uma empresa, qualquer que seja seu setor de atuação ou localização, é de fundamental interesse em estudos longitudinais e transversais na economia. Foi construída uma *proxy* para a idade da empresa a partir de um algoritmo que realiza o acompanhamento longitudinal dos tempos de emprego máximo das empresas, uma vez que as bases de dados mais utilizadas nas pesquisas brasileiras não possuem expressamente essa informação. A comparação deu-se por meio de análise amostral complexa, utilizando informações da Relação Anual de Informações Sociais (Rais) e aquelas divulgadas no sítio eletrônico da Receita Federal. Os resultados mostraram uma alta correlação entre a idade real da empresa e o tempo de emprego máximo para ela no ano de 2004, indicando tratar-se de uma *proxy* eficiente daquela idade desde que incorporadas às informações longitudinais.

ABSTRACT

The importance given to the time of performance of a company, for any economic sector and localization, is of extreme interest in longitudinal and transversal economic studies. A proxy variable for the real company's age was developed from an algorithm that was constructed based on the longitudinal accompaniment of the maximum times of job of the companies, since the most used databases in Brazilian studies do not possess strictly this kind of information. The comparison was made by applying the complex survey techniques using the information from the Annual Relation Social Information (Rais) and the brazilian Internal Revenue Service. The results had shown high correlation between the real age of the company and the time of maximum job for the year of 2004, indicating to be an efficient proxy variable for the real age, since it incorporates the longitudinal information.

1 INTRODUÇÃO

A idade das firmas possui grande importância na literatura econômica, sendo freqüentemente encontrada como medida de maturidade das empresas, histórico de sucesso empresarial ou, ainda, acumulação de conhecimento. Em virtude da indisponibilidade dessa informação em bases de dados brasileiras, tem-se utilizado o tempo de emprego máximo dos funcionários como uma *proxy* para o tempo de atuação dela no mercado. Existe uma forte correlação entre o tempo de emprego máximo de uma empresa e a sua idade; no entanto, a rotatividade das empresas brasileiras compromete a utilização dessa *proxy* para modelos de dados em painel, uma vez que as empresas podem se tornar mais jovens de um ano para o outro (i.e., saída de um empregado muito antigo da empresa). Acredita-se que, por um lado, em modelos de dados em painel, o tempo de emprego máximo acrescente uma variabilidade indesejada em razão dessa rotatividade. Por outro lado, nos modelos *cross-section* essa limitação é residual, pois esses não sofrem a influência da rotatividade ao considerarem um único período de tempo.

No presente estudo, utiliza-se a análise de dados amostrais complexos para verificar o grau de associação entre a idade de uma firma e o tempo de emprego do funcionário que nela trabalha há mais tempo. Pretende-se ainda propor uma correção para essa *proxy* a partir de um acompanhamento longitudinal da permanência dos funcionários nas firmas a partir dos dados da Relação Anual de Informações Sociais (Rais), caso a diferença entre a *proxy* e a idade real seja significante. O planejamento amostral incorporou a localização das empresas nas cinco grandes regiões brasileiras; o setor de atividade econômica cuja empresa está inserta, dividida em seis níveis, a saber: agropecuária, indústria, construção civil, serviços, comércio e governo; e o porte da empresa.

2 REVISÃO BIBLIOGRÁFICA

2.1 SIGNIFICADO ECONÔMICO DO TEMPO DA ATUAÇÃO DA EMPRESA

Encontram-se várias utilizações da idade da empresa na literatura econômica, seja no estudo dos determinantes do comércio exterior, seja no crescimento das empresas, seja no da demografia de firmas. Dentro da literatura de comércio exterior, Araújo (2005) considera o passado empresarial um determinante importante das exportações, uma vez que a competitividade passada e o sucesso empresarial podem influenciar a inserção externa das firmas, ainda que a relação entre exportação e idade não seja necessariamente linear. Especialmente em países com ambiente macroeconômico instável, a idade da empresa é utilizada para captação do histórico de sucesso no passado, baseando-se na relação da eficiência e da probabilidade de sobrevivência da empresa.

Ainda na literatura de comércio internacional, Lefebvre e Lefebvre (2000) levantam um possível conflito entre o tempo de atuação da empresa e a exportação de bens e serviços. Por um lado, firmas antigas podem ter acumulado mais conhecimento, o que possibilita construir um conjunto de capacitações que lhes permitam penetrar mais facilmente no mercado externo (BALDWIN; RAFIQUZZAMAN, 1998). Essa maturidade, contudo, pode significar resistência na busca de novos mercados e, assim, firmas mais jovens muitas vezes apresentam uma

postura mais agressiva (LEONARD-BARTON, 1992). Esses fatos mostram a necessidade de se utilizar a idade da empresa como variável de controle na verificação do relacionamento entre exportações e capacitações das firmas.

Najberg e Puga (2002) analisam a taxa de sobrevivência e a criação de postos de trabalho para empresas brasileiras no período de 1995-2000. Particular atenção é dada à grande rotatividade de empresas de pequeno porte, que respondem pela grande maioria das unidades criadas a cada ano. Os autores procuraram responder a questões, como: qual é o ciclo de vida das firmas no país e quando foram criadas as unidades hoje em atividade? A dinâmica de sobrevivência das firmas se difere por porte e/ou por setor de atividade? As firmas de menor porte são realmente as principais criadoras de emprego? Para chegarem aos resultados, os autores utilizaram a Rais, na qual a taxa de sobrevivência é calculada a partir do acompanhamento das empresas existentes no ano de 1995.

A partir da regularidade empírica verificada pela identificação de uma distribuição assimétrica para o tamanho de empresas francesas no início do século XX, Gibrat estabeleceu que a taxa de crescimento de uma empresa não depende do seu tamanho, o que ficou conhecido na literatura como a Lei de Gibrat (SUTON, 1997). Na verificação da validade empírica dessa lei, Evans (1987), Hall (1987) e Ribeiro (2002) consideram a utilização da idade das empresas como um importante determinante de controle. Ribeiro (2002) rejeita a validade de Lei de Gibrat para as empresas industriais brasileiras. Sua análise utilizou as informações da Pesquisa Industrial Anual (PIA) do Instituto Brasileiro de Geografia e Estatística (IBGE) para o período de 1996 a 1999. O autor observa que empresas novas e antigas apresentam padrões de comportamento distintos na relação entre a taxa de crescimento e o tamanho, e propõe a utilização da idade da empresa como variável instrumental no modelo de crescimento. Evans (1987) e Hall (1987) mostram que, para a indústria dos EUA, as pequenas empresas crescem mais rápido do que as grandes e concluem, ainda, que a taxa de variação do emprego depende da idade das empresas de modo inverso.

2.2 A ANÁLISE DOS DADOS CONSIDERANDO-SE O PLANO AMOSTRAL

Neste trabalho, utilizou-se um plano amostral que desagrega a população em setores de atividade econômica, região geográfica e porte da empresa. Em amostras obtidas a partir de probabilidades desiguais, pressupor na análise a existência de uma amostra aleatória simples, ou seja, independência das unidades, leva a graves erros de interpretação quando da análise da variabilidade incorporada aos parâmetros estimados. Nesses casos, inferências que se estendam à população de referência devem ser necessariamente baseadas no uso de informações sobre o desenho amostral e dos pesos amostrais, originados nos diferentes estágios de seleção e dependentes das probabilidades de inclusão de elementos na amostra.

Levar em consideração os pesos amostrais, além do plano amostral, permite uma calibração das estimativas que se pretende remeter à população.

O erro-padrão, advindo das estimativas geradas nos cálculos, deve levar em consideração o plano amostral. Caso isso não seja feito, vieses podem ser gerados subestimando-se as estimativas, visto que o tamanho amostral será maior do que na verdade deveria ser. Isso deve ser considerado, pois parte do trabalho se baseia em

testes estatísticos paramétricos de diferenças, muito sensíveis à variabilidade em cada um dos estratos.

Para que os totais da população possam ser estimados, foram utilizados resultados baseados no plano amostral em consideração. A referência utilizada para esta seção foi Pessoa e Silva (1998).

Para estimar-se os totais, levou-se em conta o vetor de totais,

$$Y = \sum_{i \in S} y_i \ , \tag{1}$$

das Z variáveis que serão estudadas na população, a partir de uma amostra. O estimador π-ponderado de Horvitz-Thompson (SÄRNDAL; SWENSSON; WRETMAN, 1992),[1] baseia-se somente nos valores coletados na amostra, e este é dado por

$$\hat{Y}_\pi = \sum_{i \in S} y_i / \pi_i \ . \tag{2}$$

Dado que as propriedades dos estimadores baseados em planos amostrais são testadas com o auxílio da distribuição de aleatorização, serão utilizadas, a partir da distribuição de probabilidade $p(s)$, a esperança (aleatorização) e a variância (aleatorização) de p como $E_p(\cdot) \, e \, V_p(\cdot)$, respectivamente.

A variância de aleatorização, por sua vez, pode ser calculada de duas formas, bem como seus estimadores. Para a primeira variância,

$$V_p\left(\hat{Y}_\pi\right) = \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \cdot \frac{y_i}{\pi_i} \cdot \frac{y_j^{'}}{\pi_j}, \tag{3}$$

este é o seu estimador:[2]

$$\hat{V}_p\left(\hat{Y}_\pi\right) = \sum_{i \in S} \sum_{j \in S} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) \cdot \frac{y_i}{\pi_i} \cdot \frac{y_j^0}{\pi_j} . \tag{4}$$

Para a segunda variância, no entanto,

$$V_p\left(\hat{Y}_\pi\right) = -0{,}5 \cdot \left[\sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \cdot \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right) \cdot \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^0 \right], \tag{5}$$

este é o estimador que lhe compete,

$$V_{SYG}\left(\hat{Y}_\pi\right) = -0{,}5 \cdot \left[\sum_{i \in U} \sum_{j \in U} \left(\frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \right) \cdot \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right) \cdot \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^0 \right] . \tag{6}$$

---

1. O estimador $\pi -$ ponderado $\hat{Y}_\pi$ é não-viciado para o total Y com respeito à distribuição de aleatorização, isto é, $\hat{Y}_\pi = \sum_{i \in S} y_i / \pi_i$ (PESSOA e SILVA, 1998).

2. O primeiro estimador de variância é não-viciado da variância de aleatorização de $\hat{Y}_\pi$, i.e., $E_p\left[\hat{V}_p\left(\hat{Y}_\pi\right)\right] = \left[V_p\left(\hat{Y}_\pi\right)\right]$.

O que pode ser levado em consideração na comparação entre as variâncias reais e seus estimadores é que mesmo com as expressões iniciais de variância de aleatorização coincidindo, pelo fato de terem tamanhos fixos, o mesmo não pode ser dito para seus estimadores. Ambos os estimadores, no entanto, são não-viciados para suas variâncias.

2.3 O EFEITO DA SELEÇÃO AMOSTRAL NA ANÁLISE DE REGRESSÃO

Kish (1965) propôs uma medida denominada Efeito do Plano Amostral (EPA), ou *Design Effect* (Deff). Seu objetivo inicial era construir uma ferramenta de planejamento de pesquisa se baseando na eficiência de diferentes planos amostrais. Posteriormente, Skinner, Holt e Smith (1989) formularam o EPA Ampliado, que compara a variância estimada da variável do desenho com a sua distribuição. A diferença conceitual entre o EPA Ampliado de Skinner e o de Kish é que o primeiro avalia o erro cometido ao se ignorar o plano amostral complexo e ao analisar os dados como se fossem extraídos por meio de uma amostra aleatória simples, enquanto o segundo é utilizado como uma ferramenta de planejamento prévio da pesquisa.

O EPA de Kish é definido como uma razão de variâncias de um estimador, calculado a partir de dois planos amostrais distintos. O efeito do plano amostral de Kish para um determinado estimador $\hat{\theta}$ é definido como sendo:

$$EPA\left(\hat{\theta}\right) = \frac{V_{PAC}\left(\hat{\theta}\right)}{V_{AAS}\left(\hat{\theta}\right)}. \tag{7}$$

Essa definição é naturalmente uma medida de eficiência relativa do plano amostral utilizado em relação a uma amostra aleatória simples, e uma das formas de se verificar a eficiência de um determinado plano amostral complexo, quando as informações populacionais não estão disponíveis.

O EPA de Kish permite antecipar o impacto do uso de esquemas amostrais alternativos sobre a precisão de estimadores de totais e médias, procurando responder a pergunta: qual é o desenho de amostragem mais eficiente?

Uma vez que uma amostra complexa foi realizada, qual é o erro cometido em se realizar uma análise ingênua? Para responder a essa questão, Skinner, Holt e Smith (1989) formularam a utilização do EPA para verificar o erro cometido ao analisar-se os dados, ignorando-se o plano amostral. Considerando-se uma estimativa consistente de variância, porém, obtida por meio de uma análise ingênua[3] denotada por $v_0 = \hat{V}_{IID}\left(\hat{\theta}\right)$. A estimativa de variância, v_0, afasta-se da variância de $\hat{\theta}$ sob o plano amostral ou modelo populacional verdadeiro[4] representado por $\hat{V}_{PAC}\left(\hat{\theta}\right)$. Para avaliar esse afastamento, localiza-se a variância obtida sob o plano amostral, $\hat{V}_{PAC}\left(\hat{\theta}\right)$, em relação à verdadeira distribuição de referência de $v_0 = \hat{V}_{IID}\left(\hat{\theta}\right)$. Como simplificação,

3. Na literatura de amostragem, a expressão "análise ingênua" refere-se a ignorar o plano amostral e a analisar os dados sob a hipótese de uma amostra aleatória simples (PESSOA e SILVA, 1998).

4. O modelo de superpopulação especifica, antes da seleção da amostra, algumas das características particulares da população, como distribuição dos dados e estratificação natural (PESSOA e SILVA, 1998). Como simplificação, considerou-se a mesma notação para a variância sob o modelo populacional e sob o plano amostral complexo – $\hat{V}_{PAC}\left(\hat{\theta}\right) = \hat{V}_{POP}\left(\hat{\theta}\right)$

no lugar da distribuição de probabilidade, utiliza-se uma medida de tendência central da distribuição de v_0, ou seja, $E(v_0)$.

O efeito de especificação incorreta do plano amostral sobre a estimativa v_0 da variância do estimador $\hat{\theta}$ é dado por:

$$EPA(\hat{\theta}, v_0) = \frac{V_{PAC}(\hat{\theta})}{E(v_0)}. \quad (8)$$

Nascimento e Moura (1990) estimaram o Efeito do Plano Amostral Ampliado para regiões metropolitanas brasileiras utilizando informações socioeconômicas da pesquisa de amostragem, realizada juntamente com o censo de 1980. Os autores obtiveram grande variação do EPA, com amplitude acima de 100 unidades. Tais resultados mostram que a variância de uma estimativa pode ser subestimada em até 100 vezes ao se ignorar, na análise, o verdadeiro delineamento amostral.

Segundo Pessoa e Silva (1998), em geral, são esperadas as seguintes conseqüências sobre o EPA ao se ignorar o plano amostral: *i*) ignorar os pesos em v_0 pode inflacionar o EPA; *ii*) ignorar a conglomeração em v_0 pode inflacionar o EPA; e *iii*) ignorar a estratificação em v_0 pode reduzir o EPA.

3 METODOLOGIA

3.1 FONTES DE INFORMAÇÃO SOBRE A IDADE DAS EMPRESAS

A informação sobre a data de abertura das firmas encontra-se disponível no Cadastro Nacional de Pessoa Jurídica do Ministério da Fazenda (CNPJ/MF) e ainda no Cadastro Central de Empresas do Instituto Brasileiro de Geografia e Estatística (Cempre/IBGE). É possível a consulta individual da data de fundação de uma empresa a partir da página eletrônica da Receita Federal.[5] O tempo de emprego máximo dos funcionários pode ser calculado por meio da Relação Anual de Informações Sociais do Ministério do Trabalho e Emprego (Rais/MTE). Ressalta-se o fato de o Cempre ser o universo das empresas brasileiras, a partir do qual são construídas as diversas amostras para as pesquisas com empresas do IBGE (PIA, Pintec, PAS, Paic, PAC, etc.), sendo atualizada continuamente a partir da compilação das informações da Receita Federal, das próprias pesquisas do IBGE e da Relação Anual de Informações Sociais. Por sua cobertura universal, o Cempre constitui-se, também, como uma importante fonte de informação das empresas, mas possui limitações de acesso relacionadas ao sigilo e à proteção desses dados. O IBGE disponibiliza, por meio das suas publicações, somente informações agregadas por setor de atividade econômica, porte de empresa e município.

3.2 O PLANO AMOSTRAL

O objetivo principal desse plano amostral é produzir estimativas para a idade real das empresas brasileiras constantes da Rais.

5. O sítio é: www.receita.fazenda.gov.br/PessoaJuridica/CNPJ/cnpjreva/Cnpjreva_Solicitacao.asp.

O plano amostral utilizou uma amostra probabilística e estratificada dos CNPJs levando-se em consideração estratos independentes das cinco regiões brasileiras e dos seis setores da economia identificados pela Classificação Nacional de Atividade Econômica (Cnae) (Agropecuária [1, 2 e 5]; Indústria [10 a 37, 40 e 41]; Construção [45], Comércio [50 a 52]; Serviços [55, 60, 61, 62, 64, 65, 66, 70 a 74, 80, 85, 90 a 93, 95, 99]; e Governo [Cnaes restantes]). Como variável de estratificação implícita, utilizou-se o pessoal ocupado, o que minimizou o tamanho da amostra sem perda de representabilidade A estratificação implícita foi realizada mediante seleção das empresas por meio de uma amostragem sistemática com probabilidade proporcional ao total de empregados da Rais.

Para o cálculo do tamanho da amostra foi utilizada a variável formada a partir dos tempos máximos de emprego das empresas que constavam da base da Rais de 2004; e estavam pelo menos uma única vez entre os anos 1993 e 2004; e possuíam pelo menos dois empregados nos anos que constavam na base de dados. Essa última restrição foi incorporada para que se evitasse a entrada de unidades que representassem apenas a pessoa física. A fórmula utilizada para o cálculo da amostra foi:

$$n_e = \frac{N_a^2 \cdot S_a^2}{(CV^2 \cdot Z_a^2) + (N_a^2 \cdot S_a^2)}, \tag{9}$$

em que:

N_a – é o total de empresas no estrato amostrado a.

S_a^2 – é a variância do TEMX no estrato amostrado a.

Z_a – é o total populacional do TEMX no estrato amostrado a.

CV – é o coeficiente de variação prefixado para o estimador do TEMX em cada estrato.

O tamanho da amostra foi calculado para cada estrato, permitindo que estimador do total da idade da empresa para todos os estratos tivesse um coeficiente de variação de 9,5%. Inicialmente se tentou utilizar um CV de 10%; no entanto, por motivos de reposição da amostra, optou-se por diminuir essa medida em detrimento de uma mudança de cunho amostral no trabalho.

O algoritmo criado levou em consideração informações dos anos de 1993 a 2004, período em que se acredita que as informações da Rais sejam confiáveis. Inicialmente, foi calculado o tempo de emprego máximo, identificado pela variável, $TEMX_{ij}$, que denota o tempo de emprego máximo da empresa i num dado ano j. A regra de cálculo TEMX de uma determinada empresa levou em consideração a diferença entre o tempo transcorrido de um período j até o ano de 2004, acrescido do respectivo $TEMX_{ij}$. Por exemplo: para uma empresa A que possuía um $TEMX_{i,1996}$ no ano de 1996 de 45 anos, na verdade seu valor da $TEMX_i$ baseado em 1996 era de 54 anos ([2004-1996]+45)+1. Da mesma forma, esse cálculo foi realizado para todos os anos da mesma empresa e o $TEMX_i$ foi obtido por meio da seguinte equação:

$$TEMX_i = máx\left(\left\lfloor 2004 - ANO_j \right\rfloor + TEMX_{ij}\right) + 1 . \tag{10}$$

Assim, para o cálculo do tamanho amostral de cada um dos estratos foi utilizada a variável $TEMX_i$. A fórmula descrita em (9) incorpora informações sobre a variabilidade e a magnitude daquela variável em cada um dos estratos, por meio dos coeficientes de variação (CV) e do total Z, respectivamente. Com isso, a inferência baseada nas unidades amostrais pôde ser feita ao nível desses estratos.

Deve ser lembrado que na amostra considerada foi utilizada a informação do CNPJ a oito dígitos, ou seja, a amostragem foi realizada em relação à sede, uma vez que a informação obtida de TEMX se baseou em todos os empregados das filiais de uma mesma empresa.

Dada a dificuldade atual de se obter informações sobre a data de fundação da empresa em bases de dados disponíveis, tais como, o Cempre (IBGE), ou a base de informações da Receita Federal, utilizou-se um método amostral. Esse método só pôde ser implementado pelo fato de existir uma forma de obtenção da informação por meios lícitos que pudesse dar algum respaldo confiável à informação. Como apresentado no início da seção, a identificação só pôde ser feita no sítio da Receita Federal. Ao todo são disponibilizados, no sítio, 15[6] campos; contudo, apenas dois foram coletados quando da obtenção das informações, as quais são a data de abertura da empresa e a Classificação Nacional de Atividade Econômica da empresa. Essa última foi pesquisada para que se pudesse responder o terceiro objetivo deste trabalho, ou seja, comparar a informação de Cnae nas duas bases de dados.[7] Desse modo, foram justapostas às informações de data de fundação da empresa, o $TEMX_i$ oriundo do algoritmo apresentado e também os $TEMX_{2004}$.

Os *softwares* utilizados para as estimações e as regressões foram o SAS versão 8.2, o Sudaan e o Stata 9.0. O primeiro foi utilizado para a montagem das bases de dados e cálculos iniciais, tais como o tamanho da amostra e a $TEMX_i$. O Sudaan, programa capaz de incorporar o plano amostral em seus cálculos, foi desenvolvido pelo Research Triangle Institute e possui várias opções de planos amostrais: estratificados, conglomerados em um ou múltiplos estágios, podendo ou não considerar a seleção das unidades com probabilidades iguais. No processo de estimação de variâncias, pode-se optar pela Linearização de Taylor, o método de Jackknife e o Balance Repeated Replication (BRR).

4 RESULTADOS

Os resultados obtidos para todo o estudo consideraram em suas estimações os efeitos que o plano amostral, ou seja, a estratificação adotada, poderia causar, quando da análise de tais efeitos. Dessa forma, estão dispostas, a seguir, nas tabelas 1, 2 e 3, as informações para o tempo máximo de emprego, medidas exclusivamente para o ano de 2004 ($TEMX_{2004}$), para os anos de 1993 a 2004 ($TEMX_i$), e também as médias ponderadas pelo peso e pelo plano amostral da idade da empresa, considerando-se as informações de data de abertura obtidas na Receita Federal (Idade RF). A estatística t,

6. (1) Número de inscrição; (2) data de abertura; (3) nome empresarial; (4) nome fantasia; (5) código de descrição da atividade econômica principal; (6) código e descrição da natureza jurídica; (7) logradouro; (8) CEP; (9) bairro/distrito; (10) município; (11) unidade da Federação; (12) situação cadastral; (13) data da situação cadastral; (14) situação especial; e (15) data da situação especial.

7. Acredita-se que existam diferenças entre as Cnaes originárias das duas bases por motivos da auto-declaração da Rais.

o p-valor e o coeficiente de variação dizem respeito a essa variável. Ela indica se há diferença estatisticamente significativa entre a variável $TEMX_i$ e a Idade RF.

Na tabela 1, pode-se verificar que não há indícios de tendência em se sub ou superestimar a idade real da empresa, baseando-se nas informações da $TEMX_i$. Os setores da economia de Agropecuária, Governo e Indústria tiveram um $TEMX_i$ médio maior do que a Idade RF; no entanto, a única diferença significativa encontrada foi no setor da construção civil.

Constata-se ainda a grande diferença média entre o $TEMX_i$ e o $TEMX_{2004}$. O primeiro é em média aproximadamente duas vezes maior do que o segundo, tanto para os setores da economia quanto para as regiões brasileiras (tabela 2). Quando se faz a comparação a que essa última tabela se refere, entretanto, a única diferença significativa se dá na Região Sul, na qual, a idade média encontrada nos dados da Receita Federal é maior.

TABELA 1

Comparação entre os parâmetros da idade da empresa somente calculada em 2004 para os anos entre 1993 e 2004 e as estimativas da idade da empresa realizadas a partir da data de abertura da Receita Federal para os setores da economia

Setor econômico	$TEMX_{2004}$	$TEMX_i$	Idade RF	Estatística t[1]	P-valor	Coeficiente de variação
Agropecuária	9,59	16,47	14,52	-1,74	0,0829	0,09
Comércio	4,4	9,97	11,55	1,75	0,0813	0,08
Construção	4,76	14,34	14,94	2,43	0,0154	0,01
Governo	15,83	22,18	17,02	-1,16	0,2467	0,07
Indústria	5,81	11,41	10,61	-1,09	0,277	0,1
Serviço	6,92	12,19	13,59	0,69	0,4932	0,17

Fonte: Rais 1993 a 2004 – Ministério do Trabalho; Receita Federal – sítio em abril de 2006.

Nota: [1] Os testes estatísticos entre $TEMX_i$ e Idade RF foram feitos considerando-se o peso e o plano mostral. Os testes estatísticos entre $TEMX_{2004}$ e Idade RF não mostaram diferenças estatisticamente significativas apenas para o governo. O procedimento *surveymeans* foi utilizado nesses cálculos.

TABELA 2

Comparação entre os parâmetros da idade da empresa somente calculada em 2004 para os anos entre 1993 e 2004 e as estimativas da idade da empresa realizadas a partir da data de abertura da Receita Federal para as regiões brasileiras

Região	$TEMX_{2004}$	TEMX	Idade RF	Estatística t[1]	P-valor	Coeficiente de variação
Centro-Oeste	4,51	9,9	10,77	1,03	0,3044	0,09
Nordeste	4,69	8,51	8,63	-1,48	0,1395	0,15
Norte	4,99	10,79	16,9	1,58	0,1149	0,26
Sudeste	6,07	12,24	13,33	1,08	0,2802	0,11
Sul	5,91	12,87	13,34	2,48	0,0137	0,07

Fonte: Rais 1993 a 2004 – Ministério do Trabalho; Receita Federal – sítio da internet em abril de 2006.

Nota: [1] Os testes estatísticos entre TEMXi e Idade RF foram feitos considerando-se o peso e o plano amostral. Os testes estatísticos entre $TEMX_{2004}$ e Idade RF mostaram diferenças estatisticamente significativas para todas as regiões. O procedimento *surveymeans* foi utilizado nesses cálculos.

Vale lembrar que os coeficientes de variação (CV) – ou seja, as dispersões relativas à média, das estimativas calculadas para a Idade RF – não ficaram muito distantes em relação ao CV de 10% utilizado no cálculo do tamanho amostral para cada um dos estratos. Isso indica que a amostra coletada condisse com o plano amostral adotado.

O coeficiente de correlação de Pearson para $TEMX_i$ e a idade real da empresa, considerando o peso e o plano amostral, foi de 0,72124. Desconsiderando-se o desenho e os pesos amostrais, essa correlação passa a ser de 0,69274. Para a variável $TEMX_{2004}$, aquela correlação foi de 0,5828. Os resultados sugerem ganhos de robustez

da idade que incorpora as informações passadas da empresa em relação à informação de um único ano.

A tabela 3 desagrega as informações das tabelas 1 e 2, permitindo que se identifique a informação cruzada entre região e setor da economia, ou seja, a classificação adotada para a obtenção dos dados.

TABELA 3

Comparação entre os parâmetros da idade da empresa somente calculada em 2004 para os anos entre 1993 e 2004 e as estimativas da idade da empresa realizadas a partir da data de abertura da Receita Federal para os setores da economia e regiões brasileiras

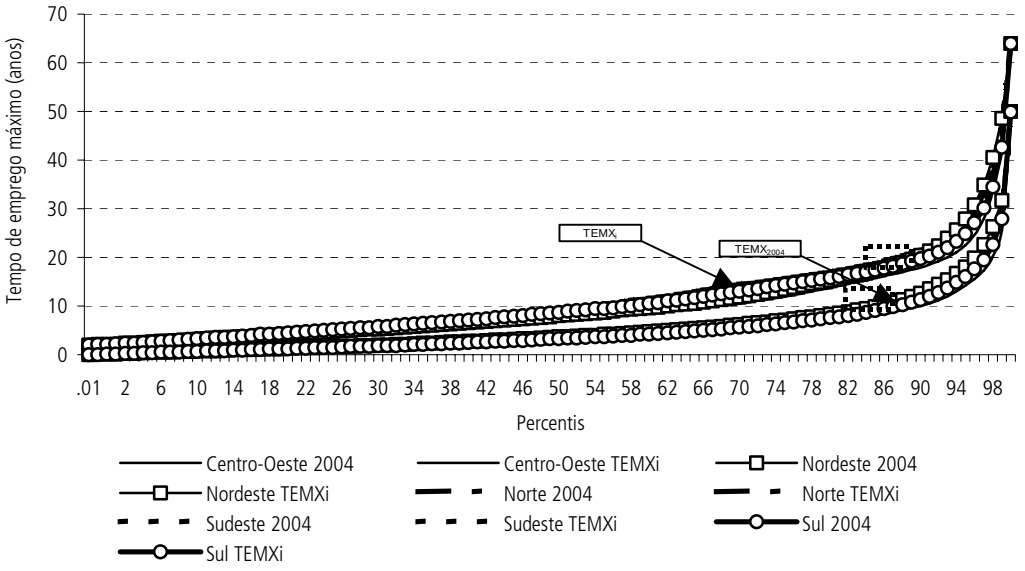
Setor econômico	Variáveis/medidas	Centro-Oeste (N=134.639)	Nordeste (N=444.726)	Norte (N=93.103)	Sudeste (N=849.128)	Sul (N=450.353)
Agropecuária (N=17.636)	N	1.592	2.272	2.601	3.383	7.708
	$TEMX_{2004}$ (m)	8,03	9,07	6,57	10,98	8,65
	$TEMX_i$ (m)	15,7	16,24	13,13	19,15	15,73
	Idade RF	11,36	9,99	19,37	19,51	12,7
	Estatística t[1]	-3,27	-1,59	3,33	0,11	-1,81
	P-valor	0,002	0,1189	0,017	0,9136	0,0762
	Coeficiente de variação	0,11	0,39	0,09	0,16	0,13
Comércio (N=935.291)	N	66.967	145.092	52.844	413.023	257.365
	$TEMX_{2004}$ (m)	3,65	4,33	3,84	4,61	4,21
	$TEMX_i$ (m)	8,54	8,98	8,59	10,23	9,75
	Idade RF	9,1	9,91	20,03	10,95	12,37
	Estatística t[1]	0,48	0,32	1,72	0,48	1,89
	P-valor	0,6351	0,7507	0,0898	0,6303	0,0644
	Coeficiente de variação	0,12	0,29	0,33	0,13	0,11
Construção (N=70.366)	N	4.714	4.767	3.044	34.704	23.136
	$TEMX_{2004}$ (m)	4,34	4,53	3,87	5,16	4,41
	$TEMX_i$ (m)	10,86	10,61	9,59	11,78	10,91
	Idade RF	9,23	12,62	11,25	15,54	16,19
	Estatística t[1]	-0,99	0,85	0,77	1,46	2,15
	P-valor	0,3233	0,3998	0,4428	0,1474	0,0349
	Coeficiente de variação	0,17	0,18	0,19	0,16	0,15
Governo (N=36.764)	N	6.847	11.904	3.784	9.988	4.241
	$TEMX_{2004}$ (m)	12,04	17	14,09	9,32	10,51
	$TEMX_i$ (m)	19,23	25,61	21,42	16,06	17,81
	Idade RF	14,48	16,46	16,68	20,44	14,99
	Estatística t[1]	-3,34	-7,11	-2,14	1,72	-0,88
	P-valor	0,0014	<0,0001	0,0357	0,0901	0,3822
	Coeficiente de variação	0,09	0,07	0,13	0,12	0,21
Indústria (N=316.291)	N	17.081	123.508	7.957	114.484	53.261
	$TEMX_{2004}$ (m)	4,49	5,84	5,14	6,88	5,52
	$TEMX_i$ (m)	9,42	11,94	11,21	12,87	11,03
	Idade RF	8,84	8,82	9,46	11,06	14,57
	Estatística t[1]	-0,45	-2,66	-0,61	-1,05	2,04
	P-valor	0,6558	0,0096	0,545	0,2997	0,454
	Coeficiente de variação	0,14	0,09	0,13	0,15	0,11
Serviço (N=595.602)	N	37.438	157.183	22.873	273.546	104.562
	$TEMX_{2004}$ (m)	5,1	6,44	5,62	6,48	5,51
	$TEMX_i$ (m)	10,55	11,59	10,8	12,51	11,42
	Idade RF	14,15	6,58	12,77	17,28	14,47
	Estatística t[1]	1,29	-3,16	0,95	1,25	1,39
	P-valor	0,2009	0,0024	0,3464	0,2159	0,1702
	Coeficiente de variação	0,19	0,24	0,16	0,22	0,15

Fonte: Rais 1993 a 2004 – Ministério do Trabalho; Receita Federal – sítio da internet em abril de 2006.

Nota: [1] Os testes estatísticos foram feitos considerando-se o peso e o plano amostral. O procedimento *surveymeans* foi utilizado nesses cálculos. A $TEMX_i$ foi calculada baseando-se nos anos de 1993 a 2004 e a $TEMX_{2004}$ somente no ano de 2004.

GRÁFICO 1

Percentis de $TEMX_i$ e $TEMX_{2004}$, por região



Fonte: Rais 1993 a 2004.

Apesar de as comparações mais importantes deste trabalho serem entre as variáveis $TEMX_i$ e Idade RF, foram confrontados os percentis das variáveis $TEMX_i$ e $TEMX_{2004}$ para que se pudesse verificar ao longo de toda a distribuição como as duas se comportam. O gráfico 1 ratifica as informações apresentadas na tabela 3, de forma que não só a média dos valores de $TEMX_i$ é maior do que $TEMX_{2004}$, mas em toda a distribuição a primeira se destaca em relação à segunda para todas as regiões brasileiras. O gráfico revela a concentração etária das empresas em anos, ou seja, olhando-se para as curvas referentes à $TEMX_i$, tem-se aproximadamente 90%[8] das empresas com idade até 20 anos. Quando, no entanto, compara-se com as outras cinco curvas, o mesmo não acontece, a idade das empresas fica por volta de 13 e 15 anos.

Deve-se atentar para o fato de a diferença estatística não poder ser dita significativa apenas com o gráfico. Foi realizado o teste de bondade do ajuste de kolmogorov-Smirnov para tal. Ele consiste na comparação entre uma função de distribuição hipotética, em que se acredita que os dados podem assumir e outra função de distribuição empírica, que, nesse caso, são variáveis do estudo.

O teste estatístico consiste na comparação de uma função distribuição empírica *S(x)*, baseada numa amostra aleatória, ou seja, informações que representem a população respectiva de forma não viesada. O teste estatístico pode ser representado de três formas; no entanto, neste trabalho será assumido somente o unicaudal. Sendo $H^*(x)$ a função de distribuição hipotética e T o maior valor medido pela altura entre *S(x)* e $H^*(x)$. O teste a ser utilizado consistirá da seguinte forma:

$$G(x) = 1 - x \sum_{j=0}^{[n(1-x)]} \binom{n}{j} \left(1 - x - \frac{j}{n}\right)^{n-j} \left(x + \frac{j}{n}\right)^{j-1}, \qquad (11)$$

em que sua distribuição assintótica é dada por:

8. Quadrados pontilhados no gráfico.

$$U(x) = \lim_{n \longrightarrow \infty} G\left(\frac{x}{\sqrt{n}}\right) = 1 - e^{-2x^2}, \tag{12}$$

e a função de distribuição aproximada de T é :

$$P(T \leq x) = \left[G(x)\right]^2. \tag{13}$$

Note que $n(1-x)$ é o maior inteiro possível a ser calculado, e essa medida pode ser usada para as duas comparações unicaudais.

O interessante desse teste é que ele pode ser aplicado para comparação de duas variáveis independentes, pois se quer identificar se ambas vêm de uma mesma população. Dessa forma, fez-se a comparação entre as variáveis $TEMX_i$ e $TEMX_{2004}$ e o teste de hipóteses é dado por:

$H_0 : H(x) \geq H^*(x)$ *para todos os x entre* $-\infty$ *a* $+\infty$.

$H_1 : H(x) < H^*(x)$ *para pelo menos um x.*

Dessa forma, pode-se verificar, por meio da tabela 4, que $TEMX_i$ e $TEMX_{2004}$ são de fato diferentes ao longo de toda a distribuição.

TABELA 4

Teste de Kolmogorov-Smirnov para a comparação das funções de distribuição do tempo de emprego em 2004 e o $TEMX_i$

	Estatística D	P-valor
$TEMX_{2004}$	0.3810	<0.0001
$TEMX_i$	<0.0001	1

Fonte: Rais 1993 a 2004 – Ministério do Trabalho; Receita Federal – sítio em abril de 2006.

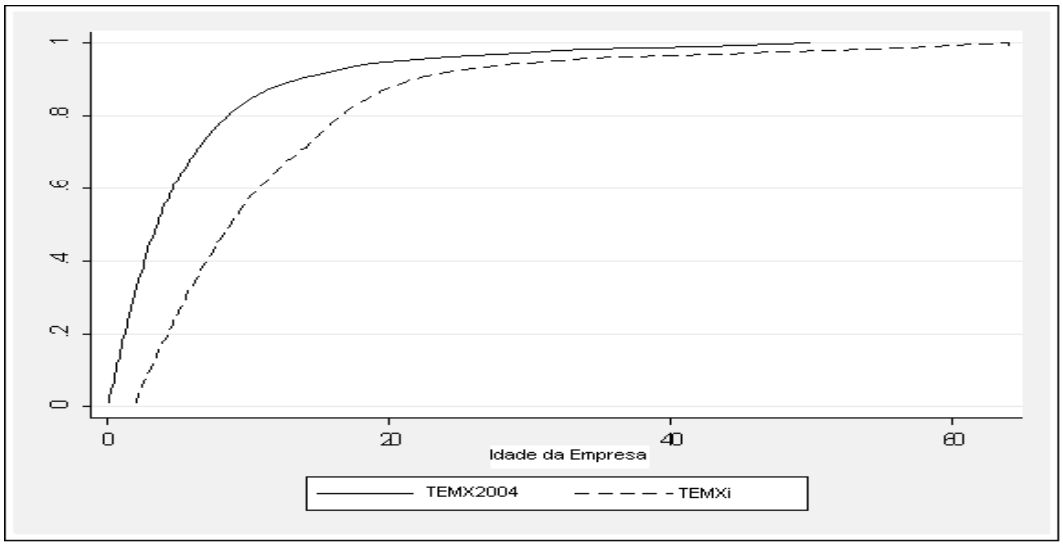
Obs.: Os testes estatísticos entre TEMXi e TEMX2004 foram realizados para toda a distribuição.

O procedimento ksmirnov do STATA 9 foi utilizado nesses cálculos.

O gráfico para as duas distribuições, por sua vez, dá-se da seguinte forma:

GRÁFICO 2

Funções de distribuição para $TEMX_i$ e $TEMX_{2004}$



Fonte: Rais 1993 a 2004.

O mesmo teste não pôde ser realizado para se comparar as funções de distribuição de $TEMX_i$ e $TEMX_{2004}$ à idade da Receita Federal, pois não se pode incorporar a esse teste o plano amostral na estimação das funções de distribuição.

Os resultados apresentados na tabela 4 são oriundos de três tipos de modelos de regressões multivariadas com as mesmas variáveis, na qual tenta-se estimar quais seriam os possíveis determinantes da idade real de uma empresa. Por falta de variáveis que pudessem indicar aquela idade em nosso cadastro, foram utilizados o número de empregados da empresa, o $TEMX_i$ e os controles de região e setor econômico na qual está inserta uma empresa por meio de efeitos fixos. Os três modelos apresentados estão divididos da seguinte forma:

a) Desconsiderando-se o peso e o plano amostral.

b) Considerando-se o peso e não o plano amostral.

c) Considerando-se o peso e o plano amostral.

Um quarto modelo foi reportado para que se pudesse mostrar a diferença entre a variável $TEMX_i$ e $TEMX_{2004}$. As comparações foram realizadas sob o modelo c, o qual inclui o peso e o plano amostral nos cálculos.

Pode-se verificar que a não utilização do peso amostral viesa consideravelmente as estimativas dos parâmetros (β) e, com isso, algumas decisões a partir dessas análises podem estar comprometidas.

A diferença dos modelos b e c não aparece nas estimativas dos parâmetros, pois esses são calculados da mesma forma, no entanto, o erro-padrão é diferente por causa da mudança na estrutura de covariância quando se inclui o plano amostral e isso pode causar danos à análise das significâncias das estimativas dos βs, como de fato acontece neste estudo.

Por motivos de baixa explicação conseguida nesse modelo, não há como predizer a variável Idade Real das Empresas pelas variáveis aqui consideradas. Foi, então, composto um novo modelo com as mesmas variáveis independentes, mas os efeitos fixos passaram a ser as Cnaes a dois dígitos e as unidades da Federação. Os primeiros modelos que apresentavam um coeficiente de determinação de aproximadamente de 0,55 passam a ter, com essa nova estruturação, um R^2 por volta de 0,74.

TABELA 5

Comparação de especificações de modelos OLS para a estimação da Idade Real da Empresa, por região e setor da economia

Variáveis	Sem peso e estimativas	Erro-padrão	P-Valor	Com peso e estimativas	Erro-padrão (Robusto)	P-Valor	Com peso e estimativas	Erro-padrão (linearizado)	P-Valor	Com peso e estimativas	Erro-padrão (linearizado)	P-Valor
Intercepto	7,9589	0,9835	<0,0001	3,4299	1,8577	0,0650	3,4299	1,8614	0,0660	8,5514	2,3035	<0,0001
$TEMX_i$	0,4796	0,0120	<0,0001	0,8560	0,0658	<0,0001	0,8560	0,0657	<0,0001	-	-	-
$TEMX_{2004}$	-	-	-	-	-	-	-	-	-	0,9796	0,0979	<0,0001
Empregados	-0,000042	0,0000	0,0010	-0,00152	0,0004	<0,0001	-0,00152	0,0004	<0,0001	-0,0011	0,0003	0,0020
Centro-Oeste	-2,0569	0,5993	0,0010	-0,3702	1,2474	0,7670	-0,3702	1,2440	0,7660	-0,4918	1,4904	0,7410
Nordeste	-1,0425	0,5999	0,0820	-1,2214	1,5178	0,4210	-1,2214	1,5167	0,4210	-1,2413	1,9548	0,5250
Norte	-1,0464	0,5921	0,0770	4,9490	4,9570	0,3180	4,9490	4,9665	0,3190	-0,6400	1,5218	0,6740
Sul	0,8982	0,6142	0,1440	-0,5103	1,1966	0,6700	-0,5103	1,2008	0,6710	0,5981	1,6616	0,7190
$Sudeste_{BASE}$	-	-	-	-	-	-	-	-	-	-	-	-
Agropecuária	-3,4805	0,7043	<0,0001	-3,2431	2,0060	0,1060	-3,2431	2,0087	0,1070	-2,1550	2,4751	0,3840
Comércio	-0,8993	0,6708	0,1800	-0,3175	1,6439	0,8470	-0,3175	1,6480	0,8470	-1,7076	2,0207	0,3980
Construção	0,2573	0,6055	0,6730	-0,6677	1,7511	0,7030	-0,6677	1,7545	0,7040	-0,5558	2,0277	0,7840

(continua)

(continuação)

Variáveis	Sem peso e estimativas	Erro-padrão	P-Valor	Com peso e estimativas	Erro-padrão (Robusto)	P-Valor	Com peso e estimativas	Erro-padrão (linearizado)	P-Valor	Com peso e estimativas	Erro-padrão (linearizado)	P-Valor
Governo	-5,0808	0,7143	<0,0001	-5,0256	2,3317	0,0310	-5,0256	2,3311	0,0310	-6,6397	2,5417	0,0090
Indústria	-0,2381	0,6258	0,7040	-2,0923	1,5223	0,1690	-2,0923	1,5258	0,1700	-2,3407	2,0088	0,2440
$Serviço_{BASE}$	-	-	-	-	-	-	-	-	-	-	-	-
Medidas	(1)	(2)	(3)	(4)								
n	1.980	1.980	1.980	1.824								
N	1.971.949	1.971.949	1.971.949	1.638.015								
Graus de liberdade	F(11,1968)	F(11,1968)	F(11,1887)	F(11,1783)								
Estatística do teste	188,44	31,33	30,75	20,35								
Prob > F	<0,00001	<0,00001	<0,00001	<0,00001								
Número de estratos	30	30	30	30								
R^2	0,513	0,5515	0,5515	0,3603								
Raiz quadrada – EQM	200,93	200,93	6,3669	7,71								

Fonte: Rais 1993 a 2004 – Ministério do Trabalho; Receita Federal – sítio da internet em abril de 2006.
Obs.: A modelagem foi realizada com pelo programa STATA 9.0.

Mesmo com essa alteração, não foram observados ganhos importantes que indicassem possibilidades de predição de valores para a Idade Estimada da Empresa. Dessa forma, considerar-se-á como variável *proxy* da Idade Real da Empresa somente a $TEMX_i$. Ela por si só já carrega bastante informação da variável-chave deste estudo.

4.1 COMPARAÇÃO DAS CLASSIFICAÇÕES NACIONAIS DE ATIVIDADES ECONÔMICAS

A outra comparação sugerida no trabalho foi a da Cnae pelas informações coletadas da Receita Federal e da Rais. Foi realizada a comparação de concordâncias entre as duas bases de origem descritas na tabela 7, e nota-se que quanto maior a desagregação menor o percentual de concordância entre as duas origens.

TABELA 6
Comparação de especificações de modelos OLS para a estimação da Idade Real da Empresa, por Cnae a dois dígitos e unidades da Federação

Variáveis	Sem peso e sem plano amostral (1)			Com peso e sem plano amostral (2)			Com peso e com plano amostral (3)		
	Estimativas	Erro padrão	P-Valor	Estimativas	Erro padrão (robusto)	P-Valor	Estimativas	Erro padrão (linearizado)	P-Valor
Intercepto	2,81107	1,34776	<0,0001	-0,50038	1,9047	0,7930	-0,5004	1,8680	0,7890
$TEMX_i$	0,47476	0,01308	0,0010	0,82874	0,0455	<0,0001	0,8287	0,0443	<0,0001
Empregados	-0,00004	0,00001	0,0370	-0,00129	0,0003	<0,0001	-0,0013	0,0003	<0,0001
Cnae 2 dígitos					Não reportado				
UF					Não reportado				
Medidas	(1)	(2)	(3)						
n	1.980	1.980	1.980						
N	1.971.949	1.971.949	1.971.949						
Graus de liberdade	F(83,1896)	F(76,1896)	F(77,1821)						
Estatística do teste	26,31	66,14	66,14						
Prob > F	<0,0001	<0,0001	<0,0001						
Número de estratos	30	30	30						
R^2	0,5353	0,7433	0,7433						
Raiz quadrada – EQM	6,17	154,87	4,9075						

Fonte: Rais 1993 a 2004 – Ministério do Trabalho; Receita Federal – sítio da internet em abril de 2006.
Obs.: A modelagem foi realizada com pelo programa STATA 9.0.

A tabela 8 mostra a maior quantidade de empresas com classificações idênticas em ambas as bases. Algumas más classificações, entretanto, foram identificadas. O setor agropecuário e o de serviços apresentam as piores classificações. Pode-se verificar que aproximadamente 14% das empresas classificadas como agropecuária na

Rais foram classificadas como comércio na Receita Federal; 9,6% como da indústria; e 17% como empresas do setor de serviços. A outra má classificação ocorreu nos setores de serviços e do governo. Aproximadamente 24% das empresas governamentais classificadas pela Rais foram classificadas como do setor de serviços na Receita Federal. Por outro lado, aproximadamente 18% das empresas classificadas pela Receita Federal como do setor governamental foram classificadas como de serviços pela Rais.

TABELA 7

Concordâncias do número de empresas entre as Classificações Nacionais de Atividades Econômicas nas bases da Receita Federal e da Relação Anual de Integração Social

Agregação	Níveis		Concordância		Total
	Rais	Receita Federal	N	%	
Setores agregados	6	6	1.873.636	95	
Cnae 2 dígitos	56	55	1.779.678	90,24	
Cnae 3 dígitos	153	162	1.700.031	86,21	**1.971.949**
Cnae 4 dígitos	286	316	1.607.183	81,5	

Fonte: Rais 1993 a 2004 – Ministério do Trabalho; Receita Federal – sítio da internet em abril de 2006.

TABELA 8

Número de empresas e efeitos do plano amostragem das informações de Cnae da Rais e da Receita Federal

Classificação Rais [(1) N;(2)%Col (3)%Lin)]	Classificação Receita Federal						
	Agropecuária	Comércio	Construção	Governo	Indústria	Serviço	**Total**
	9941	2.495	349	54	1.700	3.097	
Agropecuária	0,06	0,28	0,52	0,14	0,51	0,50	**17.636**
	0,00	14,15	1,98	0,31	9,64	17,56	
	6	886.706	19	861	16.188	31.511	
Comércio	0,06	98,44	0,03	2,24	4,89	5,04	**935.291**
	0,00	94,81	0,00	0,09	1,73	3,37	
	3	203	66.336	767	229	2.828	
Construção	0,03	0,02	99,45	1,99	0,07	0,45	**70.366**
	0,00	0,29	94,27	1,09	0,33	4,02	
				27.995	6	8.763	
Governo	0,00	0,00	0,00	72,68	0,00	1,40	**36.764**
	0,00	0,00	0,00	76,15	0,02	23,84	
	96	9.727		1.929	304.126	413	
Indústria	0,96	1,08	0,00	5,01	91,95	0,07	**316.291**
	0,03	3,08	0,00	0,61	96,15	0,13	
		1.654		6.913	8.502	578.532	
Serviço	0,00	0,18	0,00	17,95	2,57	92,54	**595.601**
	0,00	0,28	0,00	1,16	1,43	97,13	
Total	**10.046**	**900.785**	**66.704**	**38.519**	**330.751**	**625.144**	**1.971.949**

Fonte: Rais 1993 a 2004 – Ministério do Trabalho; Receita Federal – sítio da internet em abril de 2006.

Obs.: Os testes estatísticos foram feitos considerando-se o peso e o plano amostral. Os procedimentos para cálculo de EPA foram feitos no STATA 9.0.

5 CONCLUSÃO

Este trabalho investigou a consistência da utilização do tempo de emprego máximo dentro de uma empresa como estimativa para a idade da firma. Essa comparação foi possível por meio da elaboração de um plano de amostragem estratificado e da coleta da idade real das empresas por meio da internet. Observou-se que o tempo de emprego máximo (TEMX) entre os trabalhadores possui uma alta correlação com a sua idade real; porém esse precisa ser corrigido incorporando-se as informações passadas sobre o TEMX entre os trabalhadores, em anos anteriores. Visando ao maior detalhamento do estudo, foi feito um acompanhamento do tempo de emprego máximo entre os trabalhadores das empresas amostradas desde o ano de 1993 até 2004. As análises utilizaram os desenvolvimentos estatísticos em estimação, que consideram a probabilidade de seleção das empresas na amostra e o plano amostral utilizado.

O setor de construção civil apresentou as maiores inconsistências para a *proxy* de idade da firma em 2004, calculada a partir dos tempos de emprego máximo das empresas desde o ano de 1993 até 2004 ($TEMX_i$). Para aquele setor, a hipótese nula de igualdade entre a *proxy* de idade ($TEMX_i$) e a idade real da firma é rejeitada com um nível de 5% de significância. Para a agropecuária e o comércio, a hipótese de igualdade entre $TEMX_i$ e a idade real da firma só podem ser rejeitadas com 10% de significância. Os setores da indústria, governo e serviços apresentaram boa consistência para o tempo de emprego máximo dos empregados.

A comparação dentro de cada um desses setores por região geográfica permite observar que, no setor agropecuário, as maiores inconsistências se observam no Centro-Oeste e na Região Norte. O comércio apresentou rejeição da hipótese nula de igualdade entre as idades da empresa reais e estimadas somente na Região Sul. Para indústria e serviços, as diferenças estatisticamente significativas estão no Nordeste. O setor governamental apresentou discrepâncias no Norte e Nordeste.

A partir da consistência observada da variável TEMXi, foram construídas as variáveis longitudinais para todas as empresas que constavam na Rais em cada um dos anos. Ou seja, para se fazer um estudo longitudinal, recomenda-se utilizar essa variável, pois ela incorpora a variabilidade anual de uma mesma empresa em períodos distintos.

Com o objetivo de verificar-se a possibilidade de uma eventual imputação ou predição da idade real para todas as empresas da Rais, foram ajustados alguns modelos de regressão linear múltipla envolvendo o tempo de emprego máximo dos trabalhadores calculado desde 1993 e a idade real da empresa. Os modelos consideraram efeitos fixos para unidade da Federação e para Classificação Nacional de Atividade Econômica. As estimativas dos parâmetros foram obtidas por meio da Máxima Pseudoverossimilhança, que incorporam os efeitos do plano amostral e das probabilidades de seleção da empresa. Não se pretende, a partir desses modelos, caracterizar nenhuma relação econômica entre as variáveis, dada a possível endogeneidade existente nessa relação. Foi possível, entretanto, observar o ganho de informação pela incorporação do plano amostral na análise dos dados, bem como propor trabalhos futuros utilizando técnicas multivariadas de imputação de dados. Os resultados sugerem um forte potencial de utilização da informação do tempo máximo de emprego para a imputação estatística da idade real da firma, o que poderá ser desenvolvido em trabalhos futuros.

REFERÊNCIAS

AN, A.; WATTS, D. New SAS procedures for analysis of sample survey data. SUGI Proceedings. SAS Institute Inc., Cary, NC, 1998. Disponível em: <http://support.sas.com/rnd/app/papers/survey.pdf>. Acesso em: 1 abr. 2006.

ANDRADE, D. F.; SILVA, P. L. N.; BUSSAB, W. O. *Plano amostral para o Saeb 2001*. Brasília: Final Draft, 2001.

ARAÚJO, B. C. P. O. *Os determinantes do comércio internacional ao nível da firma:* evidências empíricas. Brasília: Ipea, 2005 (Texto para Discussão, n. 1.133).

BALDWIN, J. R.; RAFIQUZZAMAN, M. The determinants of the adoption lag for advanced manufacturing technologies. *Management of Technology, Sustainable Development and Eco-efficiency.* Edited by L. A. Lefebvre, R. Mason and T. Khalil, Amsterdam: Elsevier, 1998.

BINDER, D. A. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review,* n. 51, p. 279-292, 1983.

BOLFARINE, H.; BUSSAB, W. O. *Elementos de amostragem*. São Paulo: Edusp. Versão preliminar.

COCHRAN, W. G. *Técnicas de amostragem*. Segunda Edição. Rio de Janeiro: Ed. Fundo de Cultura, 1965.

_____. *Sampling techniques.* Third Edition. New York: John Wiley & Sons, 1977.

CONOVER, W. J. *Practical nonparametric statistics.* Third Edition. New York: John Wiley & Sons, 1999.

COX, D. R.; HINKLEY, D. V. *Theoretical statistics.* Londres: Chapman and Hall, 1974.

DOBSON, A. J. *An introduction to generalized linear models.* London: Chapman and Hall, 1996.

EVANS, D. Tests of alternative theories of firm growth. *Journal of Political Economy,* n. 95, p. 657-674, 1987.

FULLER, W. A. Regression analysis for sample survey. *Sankhya C*, n. 37, p. 117-132, 1975.

GARTHWAITE, P. H.; JOLLIFE, I. T.; JONES, B. *Statistical inference.* Nova York: Prentice Hall, 1995.

HALL, B. The relationship between firm size and firm growth in the US manufacturing sector. *Journal of Industrial Economics,* n. 35, p. 583-606, 1987.

HOLT, D. Introduction to part C. *In*: SKINNER, C. J.; HOLT, D.; SMITH, T. M. F. (Eds.). Analysis of Complex Surveys. Chichester: Wiley, p. 209-215, 1989.

HOLT, D.; SMITH, T. M. F.; WINTER P. D. Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society,* A, n. 143, p. 474-487, 1980.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). Pesquisa Anual de Serviços 2002, v. 33, Rio de Janeiro, IBGE, 2005 (Séries Relatórios Metodológicos).

KISH, L. *Survey sampling.* Nova York: Wiley, 1965.

LEFEBVRE, E.; LEFEBVRE, L. A. *SMEs, exports and job creation:* a firm-level analysis. Industry Canada Research Publications Program. Ottawa, Canada, Dec. 2000 (Occasional Paper, n. 26).

LEONARD-BARTON, D. Core capabilities and core rigidities: a paradox in product development. *Strategic Management Journal,* n. 13, Summer 1992, p. 111-26.

NAJBERG, S.; PUGA, F. P. O ciclo de vida das firmas e seu impacto no emprego: o caso brasileiro 1995-2000. *Revista do BNDES,* Rio de Janeiro, v. 9, n. 18, p. 149-162, dez. 2002.

NASCIMENTO, S. P. L. D.; MOURA, F. A. S. *Efeitos de conglomeração da malha setorial do censo demográfico 80*. Rio de Janeiro: IBGE, Diretoria de Pesquisas, 1990 (Série Textos para Discussão, n. 32).

PAULA, G. A. Modelo de regressão com apoio computacional. Instituto de Matemática e Estatística – Universidade de São Paulo. Disponível em: <http:www.ime.usp.br /~giapaula>. Acesso em: 27 abr. 2006.

PESSOA, D. G. C.; NASCIMENTO, S. P. L. D.; DUARTE, R. P. N. Análise de dados amostrais complexos. Escola Nacional de Ciências Estatísticas – Ence. Minicurso Sinape, 1998.

PESSOA, D.; SILVA, P. L. N. *Análise de dados amostrais complexos.* São Paulo: Associação Brasileira de Estatística, 1998. v. 1. 187p.

PFEFFERMANN, D.; NATHAN, G. Regression analysis of data fromcomplex samples. *Journal of the American Statistical Association,* n. 76, p. 681-689, 1981.

PFEFFERMANN, D. The role of sampling weights when modeling survey data. *International Statistical Review,* n. 61, p. 317-337, 1993.

RIBEIRO, E. P. *Distribuição e dinâmica do tamanho de empresas industriais*. Escola Nacional de Ciências Estatísticas e Universidade Federal do Rio Grande do Sul, 2002 Texto não publicado.

SKINNER, C. J.; HOLT, D.; SMITH, T. M. *Analysis of complex surveys.* New York: Wiley, 1989.

SUTTON, J. Gibrat's legacy. *Journal of Economic Literature*, vol. XXXV, Mar. 1997.

ANEXO

1 O MÉTODO DE PSEUDOMÁXIMA VEROSSIMILHANÇA

Uma das alternativas de se incorporar o plano amostral numa análise de regressão é utilizar o método da máxima pseudoverossimilhança. A estimação dos parâmetros de um modelo que incorpora o efeito do plano amostral foi considerada em Fuller (1975) e Holt, Smith e Winter (1980), Pfeffermann e Nathan (1981) Cox e Hinkley (1974), Skinner, Holt, Smith (1989), Garthwaite, Jollife, Jones (1995) e Pessoa e Silva (1998).

Ao se supor os vetores observados, $\mathbf{y_i} = (y_{i1},\ldots,y_{iR})$, das variáveis de pesquisa do elemento i gerados pelos vetores aleatórios $\mathbf{Y_i}$, para $i \in U$. Suponha ainda que na população todos os elementos são conhecidos independentemente distribuídos com densidade $f(\mathbf{y},\boldsymbol{\theta})$, em que $\theta = (\theta_1,\ldots,\theta_K)^0$ é um vetor de parâmetros desconhecidos. A verossimilhança populacional será dada por:

$$L_U(\boldsymbol{\theta}) = \sum_{i\in U} \log f(\mathbf{y_i},\boldsymbol{\theta}).$$

As equações de verossimilhança amostrais correspondentes são dadas por:

$\sum_{i\in U} \mathbf{u_i}(\boldsymbol{\theta})$, onde $\mathbf{u_i}(\boldsymbol{\theta}) = \partial/\partial\boldsymbol{\theta}[\log f(y_i,\boldsymbol{\theta})]$

é o vetor de escores.[9]

A estatística $T = \sum_{i\in U} \mathbf{u_i}(\boldsymbol{\theta})$ é a soma dos vetores de escore na população. Para estimar esse vetor de totais, pode-se utilizar um estimador linear ponderado na forma $T = \sum_{i\in U} \mathbf{w_i u_i}(\boldsymbol{\theta})$, em que w_i são os pesos amostrais. O estimador de máxima pseudoverossimilhança $\hat{\theta}_{MPV}$ de $\hat{\theta}_U$ será a solução das equações de pseudo-verossimilhança dado por $\hat{T} = \sum_{i\in S} \mathbf{w_i u_i}(\boldsymbol{\theta})$.

Por meio da linearização de Taylor e considerando os resultados de Binder (1983), podem ser obtidas a variância de aleatorização assintótica do estimador θ_{MPV} e o estimador da variância do estimador, dados respectivamente por

$$V_p(\hat{\boldsymbol{\theta}}_{\mathbf{MPV}}) \cong [J(\boldsymbol{\theta}_U)]^{-1} V_p\left[\sum_{i\in s} \mathbf{w_i u_i}(\boldsymbol{\theta})\right][J(\boldsymbol{\theta}_U)]^{-1}$$

e

$$V_p(\hat{\boldsymbol{\theta}}_{\mathbf{MPV}}) \cong [\hat{J}(\hat{\boldsymbol{\theta}}_{\mathbf{MPV}})]^{-1} \hat{V}_p\left[\sum_{i\in s} \mathbf{w_i u_i}(\hat{\boldsymbol{\theta}}_{\mathbf{MPV}})\right][\hat{J}(\hat{\boldsymbol{\theta}}_{\mathbf{MPV}})]^{-1}$$

9. Sob condições de regularidade (COX; HINKLEY, 1974) a solução desse sistema é o estimador de máxima pseudoverossimilhança de um censo. Pfeffermann (1993) define $\boldsymbol{\theta}_{\mathbf{U}}$ como sendo a Quantidade Descritiva Populacional Correspondente só calculável em um censo e se constitui o pseudoparâmetro eleito como alvo da inferência que incorpora o planejamento amostral.

$$J(\boldsymbol{\theta}_U) = \left.\frac{\partial T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_U} = \sum_{i\in U} \left.\frac{\partial \mathbf{u}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_U} \Rightarrow$$

$$J(\boldsymbol{\theta}_U) = \left.\frac{\partial \hat{T}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MPV}} = \sum_{i\in U} w_i \left.\frac{\partial \mathbf{u}_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MPV}},$$

$V_p\left[\sum_{i\in s} \mathbf{w_i u_i}\left(\hat{\boldsymbol{\theta}}_{MPV}\right)\right]$ é a matriz de variância (aleatorização) do estimador do total populacional dos escores e $\hat{V}_p\left[\sum_{i\in s} \mathbf{w_i u_i}\left(\hat{\boldsymbol{\theta}}_{MPV}\right)\right]$ é um estimador consistente para essa variância. Binder (1983) mostrou que a distribuição assintótica de $\hat{\boldsymbol{\theta}}_{MPV}$ é normal multivariada, isto é,

$$\hat{V}_p\left[\left(\hat{\boldsymbol{\theta}}_{MPV}\right)\right]^{-\frac{1}{2}}\left(\hat{\boldsymbol{\theta}}_{MPV} - \boldsymbol{\theta}_U\right) \sim NM(0, I).$$

2 QUADRADOS MÍNIMOS ORDINÁRIOS COM OS PESOS AMOSTRAIS

Uma das possibilidades de se incluir os pesos seria estimar

$$\hat{\beta}_w = \left(\sum_{i\in S} w_i \mathbf{x_i'} \mathbf{x_i}\right)^{-1} \sum_{i\in S} w_i \mathbf{x_i'} y_i = \left(\mathbf{X_s'} \mathbf{W_s} \mathbf{X_s}\right)^{-1} \mathbf{X_s'} \mathbf{W_s} \mathbf{Y_s}$$

em que $w_i = \pi_i^{-1}$, y_i e $W_s = diag\{w_i, i \in s\}$. O estimador $\hat{\beta}_w$ foi primeiramente considerado sob o contexto de uma regressão com heterocedasticidade, entretanto, Fuller (1975), Holt, Smith e Winter (1980), Pfeffermann e Nathan (1981) estudaram as suas propriedades sob um plano de estimação complexo.

2.1 Estimadores de pseudomáxima verossimilhança dos parâmetros do modelo

Considerando-se os estimadores de máxima verossimilhança da população denotados por $\mathbf{B}$ e S_e, e utilizando-se os pesos w_i para obter os estimadores de máxima pseudoverossimilhança de β e σ, as equações de pseudoverossimilhança correspondentes são descritas como

$$\sum_{i\in S} w_i \mathbf{u_i}\left(\hat{\mathbf{B}}_\mathbf{w}\right) = \sum_{i\in S} w_i \mathbf{x_i}\left(y_i - \mathbf{x_i'}\hat{\mathbf{B}}_w\right) = \mathbf{x}_S' \mathbf{W}_S \mathbf{y}_S - \left(\mathbf{x}_S' \mathbf{W}_S \mathbf{y}_S\right)\hat{\mathbf{B}}_w = \mathbf{0},$$

$$\sum_{i\in S} w_i u_\mathbf{i}\left(s_e^w\right) = \sum_{i\in S} w_i \left[\left(y_i - \mathbf{x_i'}\hat{\mathbf{B}}_w\right)^2 - s_e^w\right] = \left(\mathbf{y}_S - \mathbf{x}_S'\hat{\mathbf{B}}_S\right)' \mathbf{W_s}\left(\mathbf{y}_S - \mathbf{x}_S'\hat{\mathbf{B}}_S\right) - \left(\mathbf{1}_S' \mathbf{W_s} \mathbf{1}_S\right) s_e^w = \mathbf{0},$$

em que $\mathbf{W}_s = diag\left[\left(w_{i_1},\ldots,w_{i_n}\right)\right]$ é uma matriz com os pesos dos elementos da amostra na diagonal principal, $\hat{\mathbf{B}}_w$ e s_e^w são os estimadores de máxima pseudoverossimilhança (MPV) de β e σ_e, respectivamente.

Considerando-se a hipótese de singularidade[10] de $\mathbf{x}_s'\,\mathbf{W}_s\mathbf{y}_s$, pode-se obter os seguintes estimadores MPV do modelo:

$$\hat{\mathbf{B}}_w = \left(\mathbf{x}_S'\mathbf{W}_S\mathbf{y}_S\right)^{-1}\left(\mathbf{x}_S'\mathbf{W}_S\mathbf{y}_S\right),$$

$$s_e^w = \left(\mathbf{1}_S'\mathbf{W}_S\mathbf{1}_S\right)^{-1}\left(\mathbf{y}_S - \mathbf{x}_S\hat{\mathbf{B}}_w\right)'\mathbf{W}_S\left(\mathbf{y}_S - \mathbf{x}_S\hat{\mathbf{B}}_w\right) = \left(\mathbf{1}_S'\mathbf{W}_S\mathbf{1}_S\right)^{-1}\mathbf{y}_S'\left[\mathbf{W}_S - \mathbf{W}_S\mathbf{x}_S\left(\mathbf{x}_S'\mathbf{W}_S\mathbf{x}_S\right)\mathbf{x}_S'\mathbf{W}_S\right]\mathbf{y}_S .$$

Ao substituir-se $\mathbf{W}_s$ por $\mathbf{\Pi}_s^{-1} = diag\left(\pi_i : i \in S\right)$, obtêm-se os estimadores π-ponderados de mínimos quadrados:

$$\hat{\mathbf{B}}_w = \left(\mathbf{x}_S'\mathbf{\Pi}_S^{-1}\mathbf{y}_S\right)^{-1}\left(\mathbf{x}_S'\mathbf{\Pi}_S^{-1}\mathbf{y}_S\right) \text{ e}$$

$$s_e^w = \left(\mathbf{1}_S'\mathbf{\Pi}_S^{-1}\mathbf{1}_S\right)^{-1}\left(\mathbf{y}_S - \mathbf{x}_S\hat{\mathbf{B}}_w\right)'\mathbf{\Pi}_S^{-1}\left(\mathbf{y}_S - \mathbf{x}_S\hat{\mathbf{B}}_w\right).$$

3 EFEITO DO PLANO AMOSTRAL PARA SETORES DA ECONOMIA

TABELA 8

Número de empresas e efeitos do plano amostragem das informações de Cnae da Rais e da Receita Federal

Classificação Rais [(1) N;(2)%Col (3)%Lin)]	Classificação Receita Federal						
	Agropecuária	Comércio	Construção	Governo	Indústria	Serviço	**Total**
	9941	2.495	349	54	1.700	3.097	
Agropecuária	0,06	0,28	0,52	0,14	0,51	0,50	**17.636**
	0,00	14,15	1,98	0,31	9,64	17,56	
	6	886.706	19	861	16.188	31.511	
Comércio	0,06	98,44	0,03	2,24	4,89	5,04	**935.291**
	0,00	94,81	0,00	0,09	1,73	3,37	
	3	203	66.336	767	229	2.828	
Construção	0,03	0,02	99,45	1,99	0,07	0,45	**70.366**
	0,00	0,29	94,27	1,09	0,33	4,02	
				27.995	6	8.763	
Governo	0,00	0,00	0,00	72,68	0,00	1,40	**36.764**
	0,00	0,00	0,00	76,15	0,02	23,84	
	96	9.727		1.929	304.126	413	
Indústria	0,96	1,08	0,00	5,01	91,95	0,07	**316.291**
	0,03	3,08	0,00	0,61	96,15	0,13	
		1.654		6.913	8.502	578.532	
Serviço	0,00	0,18	0,00	17,95	2,57	92,54	**595.601**
	0,00	0,28	0,00	1,16	1,43	97,13	
Total	**10.046**	**900.785**	**66.704**	**38.519**	**330.751**	**625.144**	**1.971.949**

Fonte: Rais 1993 a 2004 – Ministério do Trabalho; Receita Federal – sítio da internet em abril de 2006.

Obs.: Os testes estatísticos foram feitos considerando-se o peso e o plano amostral. Os procedimentos para cálculo de EPA foram feitos no STATA 9.0.

10. A hipótese de singularidade de $\mathbf{x}_s'\,\mathbf{W}_s\mathbf{y}_s$ não seria satisfeita se $w_i = 0$, para $i \in S$.

EDITORIAL

Coordenação
Iranilde Rego

Supervisão
Aeromilson Mesquita

Revisão
Luísa Guimarães Lima
Maria Carla Lisboa Borba
Camila de Paula Santos (estagiária)
Karen Varella Maia Corrêa (estagiária)
Olavo Mesquita de Carvalho (estagiário)
Sheila Santos de Lima (estagiária)

Editoração
Bernar José Vieira
Elidiane Bezerra Borges
Luis Carlos da Silva Marques
Gustavo de Souza Ferraz de Oliveira
Rosa Maria Banuth Arendt

Brasília
SBS – Quadra 1 – Bloco J – Ed. BNDES, 9º andar
70076-900 – Brasília – DF
Fone: (61) 3315-5090
Fax: (61) 3315-5314
Correio eletrônico: editbsb@ipea.gov.br

Rio de Janeiro
Av. Nilo Peçanha, 50, 6º andar – Grupo 609
20044-900 – Rio de Janeiro – RJ
Fone: (21) 3515-8433
Fax: (21) 3515-8402
Correio eletrônico: editrj@ipea.gov.br

Tiragem: 130 exemplares