Presidente da República
Luíz Inácio Lula da Silva

Ministro do Planejamento, Orçamento e Gestão
Paulo Bernardo Silva

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE

Presidente
Eduardo Pereira Nunes

Diretor Executivo
Sérgio da Costa Côrtes

ÓRGÃOS ESPECÍFICOS SINGULARES

Diretoria de Pesquisas
Wasmália Socorro Barata Bivar

Diretoria de Geociências
Guido Gelli

Diretoria de Informática
Luiz Fernando Pinto Mariano

Centro de Documentação e Disseminação de Informações
David Wu Tai

Escola Nacional de Ciências Estatísticas
Pedro Luis do Nascimento Silva

Ministério do Planejamento, Orçamento e Gestão
Instituto Brasileiro de Geografia e Estatística - IBGE

REVISTA BRASILEIRA DE ESTATÍSTICA

volume 64 número 222 julho/dezembro 2003

ISSN 0034-7175

R. bras. Estat., Rio de Janeiro, v. 64, n. 222, p. 1-89, jul./dez. 2003

Instituto Brasileiro de Geografia e Estatística - IBGE
Av. Franklin Roosevelt, 166 - Centro - 20021-120 - Rio de Janeiro - RJ - Brasil

Revista Brasileira de Estatística, ISSN 0034-7175

Órgão oficial do IBGE e da Associação Brasileira de Estatística – ABE.

Publicação semestral que se destina a promover e ampliar o uso de métodos estatísticos (quantitativos) na área das ciências econômicas e sociais, através de divulgação de artigos inéditos.

Temas abordando aspectos do desenvolvimento metodológico serão aceitos, desde que relevantes para os órgãos produtores de estatísticas.

Os originais para publicação deverão ser submetidos em três vias (que não serão devolvidas) para:

Francisco Louzada-Neto
Editor responsável – RBEs – IBGE.
Av. República do Chile, 500 – Centro
20031-170 – Rio de Janeiro, RJ.

Os artigos submetidos às RBEs não devem ter sido publicados ou estar sendo considerados para publicação em outros periódicos.

A Revista não se responsabiliza pelos conceitos emitidos em matéria assinada.

Editoração
Helem Ortega da Silva - Coordenação de Métodos e Qualidade - DPE/COMEQ/IBGE

Impressão
GráficaDigital/Centro de Documentação e Disseminação de Informações - CDDI/IBGE, em 2004.

Capa
Renato J. Aguiar – Coordenação de Marketing/CDDI/IBGE

Ilustração da Capa
Marcos Balster – Coordenação de Marketing/CDDI/IBGE

Revista brasileira de estatística/IBGE, - v.1, n.1 (jan./mar.1940), - Rio de Janeiro:IBGE, 1940-

v.

Trimestral (1940-1986), semestral (1987-).

Continuação de: Revista de economia e estatística.

Índices acumulados de autor e assunto publicados no v.43 (1940-1979) e v. 50 (1980-1989).

Co-edição com a Associação Brasileira de Estatística a partir do v.58.

ISSN 0034-7175 = Revista brasileira de estatística.

I. Estatística – Periódicos. I. IBGE. II. Associação Brasileira de Estatística.

IBGE. CDDI. Div. de Biblioteca e Acervos Especiais CDU 31 (05)
RJ-IBGE/88-05 (rev.98) PERIÓDICO

Impresso no Brasil/*Printed in Brazil*

Sumário

Nota do Editor

É com grande satisfação que inicio minhas atividades com editor RBEs. Aproveito a oportunidade para agradecer a minha nomeação pela diretoria da ABE e parabenizar o editor anterior, Prof. Renato Martinz Assunção, juntamente com sua equipe de editores, pelo excelente trabalho desenvolvido durante sua gestão.

Gostaria de chamar a atenção para a nova política editorial da revista, agora explicitando que publica trabalhos relevantes em estatística aplicada, não havendo limitação no assunto ou matéria em questão. Também agradeço aos editores atuais da revista, bem como ao Prof, Pedro Nascimento e Silva pela participação ativa em todo este processo de re-definições.

Além disso, para agilizar o processo de avaliação dos artigos, optamos por transformar o processo editorial da RBEs unicamente em eletrônico. Desta forma, os artigos devem ser submetidos somente via email. Após a submissão o autor correspondente receberá um código para acompanhamento o processo de avaliação do artigo.

Neste novo número da RBEs quatro artigos de áreas de aplicações distintas são apresentados. O primeiro aborda o relacionamento entre salário e escolaridade de mulheres no Brasil tratado via regressão quantílica. O segundo apresenta uma análise de resultados da prova tuberculínica por meio de inferência Bayesiana. O terceiro artigo é direcionado a modelagem de resposta aleatória na avaliação de itens sensíveis. Finalmente, o quarto artigo aborda modelos de fronteira de produção. Uma ótima leitura.

Francisco Louzada-Neto
Editor Responsável

Escolaridade e salário das mulheres no Brasil: uma aplicação de regressão quantílica

Marinalva Cardoso Maciel*
Ana Katarina T. de N. Campêlo**
Maria Cristina Falcão Raposo***

Resumo

A educação é um dos fatores de maior peso na determinação da renda de um indivíduo e a finalidade deste artigo é analisar a evolução do comportamento do salário das mulheres em função de seu grau de instrução, na década de 1990 no Brasil. A análise baseia-se na técnica de regressão quantílica. Os resultados apontam para estimativas distintas dos coeficientes das variáveis indicadoras educacionais ao longo dos quantis, indicando diferentes respostas da variável dependente (o salário) ao nível educacional, em pontos distintos de sua distribuição condicional. O retorno em educação, conforme esperado, é de maior magnitude para os quantis mais elevados e há uma tendência de crescimento do mesmo para tais quantis no período estudado. Outro resultado que se revela interessante é o aumento do nível de exigência do mercado de trabalho relativamente ao grau de instrução da mulher. Houve uma valorização do trabalho de profissionais mais qualificados ao longo do período, ou seja, pessoas mais qualificadas passaram a ganhar mais e pessoas menos qualificadas passaram a ganhar menos.

* Endereço para correspondência: Deptº de Estatística - UFPA - Cidade Univ., Recife/PE, 50670-901 - e-mail: nalva@ufpa.br

** Deptº de Economia - UFPE, Rua Édson Álvares, no 175 Apt.1103 Casa Forte, Recife/PE, 52061-450 - e-mail: campelo@ufpe.br

*** Deptº de Estatística - UFPE, Rua Nogueira Lima, no 85 Hipódromo, Recife/PE, 52045-700 - e-mail: cristina@de.ufpe.br

1. Introdução

A crescente participação da mulher no mercado de trabalho constitui uma das mais importantes transformações socioeconômicas da atualidade e tem sido o foco de várias pesquisas em todo o mundo (ver, por exemplo, Zylberstajn et al. (1985), Santos e Sedlacek (1990), Stelcner et al. (1992), Jatobá (1993) e Jacobsen (1999)). Segundo Barros, Jatobá e Mendonça (1997) esta atenção advém da necessidade de entender às motivações e mecanismos propulsores de tal fenômeno. Dentre tais fatores encontra-se a educação. O nível educacional (medido em anos completos de estudos) tem sido apontado na literatura afim como determinante da taxa de participação no mercado de trabalho e também dos níveis salariais. Desta forma, há um interesse em se estudar esta última relação, o retorno em educação[1] para as mulheres, a qual será o objeto de análise deste artigo.

Estudos sobre a relação entre escolaridade e salário têm sido desenvolvidos desde Becker (1964) e Mincer (1974). Trabalhos semelhantes foram realizados com dados brasileiros, dentre os quais podemos citar os artigos de Leal e Verlang (1991) e Araújo (1996), os quais calculam o retorno em educação para o Brasil através da estimação de versões expandidas do modelo de Mincer[2] pelo método de mínimos quadrados ordinários (MQO). No trabalho de Araújo (1996) tais estimativas são questionadas, em decorrência da violação das suposições do modelo clássico. Para corrigir o viés, o autor usou o procedimento de Heckman em dois estágios. O artigo de Ueda e Hoffman (2002), por sua vez, usa um modelo mais geral no qual os níveis educacionais tomam a forma de variáveis indicadoras, considerando a possibilidade de haver não-linearidades no retorno em educação.

Uma das principais críticas a estudos que utilizam a técnica de MQO é a comum violação das hipóteses inerentes a este modelo e o fato de que o mesmo limita-se ao efeito locacional, fornecendo assim uma visão incompleta da relação entre as variáveis. Este fato foi reconhecido por diversos autores, destacando-se Mosteller e Tukey (1977), que admitiram que o ideal seria ir além da regressão na média, através da estimação de várias curvas de regressão referentes aos diversos percentis da distribuição, resultando em uma análise mais detalhada do conjunto de dados.

---

[1] O retorno em educação é dado por $100\,[\exp(\beta_{edu}) - 1]$, onde β_{edu} é o coeficiente da variável educação (anos completos de estudo) na equação do salário onde a variável dependente é dada pelo logaritmo do salário. Ver a Seção 2 (Métodos) para o caso em que a variável educação é expressa na forma de variáveis indicadoras.

[2] Modelo que utiliza o logaritmo natural do salário em função da educação e experiência, dentre outras variáveis explicativas.

O avanço da computação possibilitou a disseminação de técnicas de regressão mais robustas[3], as quais apresentam uma série de vantagens em comparação ao método tradicional no caso de violações das suposições do modelo clássico. A regressão quantílica, usada neste estudo, é um método semiparamétrico e faz parte desta classe de técnicas mais robustas de estimação. As propriedades da regressão quantílica proporcionam ganhos de eficiência e precisão relativamente à técnica de MQO, em especial nos casos em que os erros não têm distribuição normal ou quando a variável dependente apresenta valores extremos. Outro grande atrativo da regressão quantílica é que a mesma permite caracterizar toda a distribuição condicional da variável resposta, dadas as variáveis explicativas, fornecendo então uma idéia sobre a forma desta e indicando se há heterocedasticidade, assimetrias ou ainda múltiplas modas. Assim, a regressão quantílica possibilita ao investigador ter uma visão mais completa da relação entre as variáveis.

O artigo de Buchinsky (1994) usa a técnica de regressão quantílica para estimar o retorno em educação para a população dos Estados Unidos e conclui que o mesmo apresenta um padrão similar nas décadas de 1960, 1970 e 1980. Entretanto, os retornos em educação diferem ao longo dos quantis, sendo significativamente mais elevados para os quantis superiores. Em Buchinsky (1997), a análise foi feita para homens brancos americanos, nos anos de 1972, 1979, 1985 e 1992. Os resultados neste artigo revelam padrões distintos para o retorno em educação no período, para vários níveis de educação e experiência[4]. Ao longo dos quantis, na maioria dos casos, há diferenças significativas no retorno em educação. O caso feminino é por sua vez estudado em Buchinsky (1996) para os anos de 1968, 1973, 1979, 1986 e 1990. Os resultados mostraram que, de forma geral, os retornos em educação aumentaram sensivelmente para o grupo jovem (mulheres na faixa etária de 20 a 34 anos) ao longo do período, mas muito pouco para os grupos com mais idade. Os retornos geralmente são altos para os quantis inferiores do grupo jovem no início do período estudado, mas esta ordem se reverte nos últimos anos. Em Hartog, Pereira e Vieira (1999), foi feita uma análise semelhante por gênero com dados de Portugal, nos anos de 1982, 1986 e 1992. Neste trabalho as estimativas do retorno em educação tendem a ser altas para níveis educacionais mais elevados e o retorno para as mulheres é sempre inferior ao dos homens. De forma geral, os retornos vão aumentando do primeiro ao último quantil estudado, tanto para homens quanto para mulheres.

---

[3] Técnicas menos sensíveis a pequenos desvios das suposições básicas dos modelos.

[4] A variável experiência mede a experiência profissional, ou seja, o número de anos em que o indivíduo exerceu atividades profissionais.

O presente estudo tem como principal objetivo analisar a relação entre os salários auferidos pelas mulheres no Brasil e seu grau de escolaridade. Mais especificamente, foi aplicado o método de regressão quantílica para investigar as variações no retorno em educação para mulheres, usando os dados da Pesquisa Nacional por Amostra de Domicílio (PNAD) nos anos de 1992 e 1999. Os resultados apontam para estimativas distintas dos coeficientes das variáveis indicadoras educacionais ao longo dos quantis, indicando diferentes respostas da variável dependente (o salário) ao nível educacional em pontos distintos de sua distribuição condicional. O retorno em educação é de maior magnitude para os quantis superiores e há uma tendência de crescimento do mesmo para tais quantis no período estudado. Outro resultado que merece destaque é o aumento do nível de exigência do mercado de trabalho relativamente à formação educacional da mulher. Houve uma valorização do trabalho dos profissionais mais qualificados ao longo do período, ou seja, pessoas mais qualificadas passaram a ganhar mais e pessoas menos instruídas passaram a ganhar menos.

Além desta seção introdutória, este artigo compõe-se de mais quatro seções. Na seguinte, Seção 2, é apresentada a metodologia utilizada neste estudo. Na Seção 3 tem-se a descrição do banco de dados e estatísticas básicas (médias, desvios padrão e medianas). Na Seção 4 são apresentadas as estimativas dos modelos e discutidos os resultados. A última seção é reservada para as conclusões do trabalho.

2. Métodos

Este artigo analisa, através da técnica de regressão quantílica, o impacto da educação no salário das mulheres no Brasil em diferentes pontos da distribuição condicional deste último. O modelo também será estimado por MQO para fins comparativos. A relação entre educação e salário será analisada primeiramente na forma linear e em seguida será usado um modelo mais abrangente, no qual a variável educação é expressa na forma de variáveis indicadoras dos diversos níveis educacionais, objetivando captar a heterogeneidade desta variável e a possível presença de não-linearidades. Na subseção 2.1 será descrita em maior detalhe a técnica de regressão quantílica e, a seguir, na subseção 2.2, tem-se os modelos utilizados na aplicação.

2.1. Regressão quantílica

A técnica de regressão quantílica foi introduzida por Koenker e Basset em 1978 e pode ser vista como uma extensão dos quantis ordinários em um modelo de locação para uma classe

mais geral de modelos em que os quantis condicionais têm a forma linear. Os autores definem o θ-ésimo quantil de regressão como a solução da seguinte função objetivo:

$$\min_{\beta} \frac{1}{n} \sum_{i:y_i \geq x_i\beta} \theta|y_i - x_i\beta| + \sum_{i:y_i < x_i\beta} (1-\theta)|y_i - x_i\beta| = \min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho_\theta(y_i - x_i\beta),$$

onde ρ é a função *check* definida por:

$$\rho_\theta(z) = \begin{cases} \theta z, & z \geq 0 \\ (\theta - 1)z, & z < 0 \end{cases}$$

O modelo especifica a função quantil condicional da variável dependente y, dada a matriz de regressores X, como:

$$Q_\theta(y|X) = X\beta_\theta, \quad \theta \in [0,1]$$

De acordo com Buchinsky (1997), a regressão quantílica[5] apresenta características que a torna extremamente interessante, tais como: i) os modelos podem ser usados para caracterizar toda a distribuição condicional de uma variável resposta dado um conjunto de regressores; ii) o modelo tem uma representação na forma de programação linear que facilita a estimação dos parâmetros; iii) a função objetivo da regressão quantílica é uma soma ponderada de desvios absolutos, fornecendo uma medida de locação robusta, de modo que o vetor de coeficientes estimado não é sensível a observações extremas na variável dependente; iv) quando os erros não seguem a distribuição normal os estimadores de regressão quantílica podem ser mais eficientes que os estimadores de mínimos quadrados; v) soluções diferentes para quantis distintos podem ser interpretadas como diferenças na resposta da variável dependente às mudanças nos regressores, em vários pontos da distribuição condicional da mesma.

Para estudar o comportamento assintótico da estimativa de regressão quantílica são feitas as seguintes suposições adicionais:

i) a distribuição dos erros, F_ε, tem densidade contínua e estritamente positiva, f_ε, para todo z, tal que $0 < F_\varepsilon(z) < 1$.

ii) $\lim_{n\to\infty} n^{-1} X'X = D$, uma matriz positiva definida.

---

[5] A regressão de mínima soma de erros absolutos L_1 é um importante caso particular deste método. Para mais detalhes ver Koenker e Basset (1982), Buchinsky (1997), Koenker e Machado (1999) e Koenker e Hallock (2001).

Koenker e Bassett (1978) mostraram que, para erros independentes e identicamente distribuídos (i.i.d.), tem-se:

$$\sqrt{n}\left(\hat{\beta}_\theta - \beta_\theta\right) \xrightarrow{d} N(0, \Lambda_\theta), \quad \text{onde } \Lambda_\theta = \frac{\theta(1-\theta)}{f_\varepsilon^2\left(F_\varepsilon^{-1}(\theta)\right)} D^{-1}.$$

Dessa forma, a precisão assintótica da estimativa de regressão quantílica para erros i.i.d. depende basicamente da quantidade:

$$s(\theta) = \left[f_\varepsilon\left(F_\varepsilon^{-1}(\theta)\right)\right]^{-1},$$

usualmente chamada de função "sparsity"[6].

Para erros não i.i.d, o limite da matriz de covariância assume a forma do "Huber Sandwich" (ver Koenker e Portnoy,1997):

$$\sqrt{n}\left(\hat{\beta}_\theta - \beta_\theta\right) \xrightarrow{d} N\left(0, H_n^{-1} J_n H_n^{-1}\right), \quad \text{onde } J_n(\theta) = \theta(1-\theta) n^{-1} X'X$$

e

$$H_n(\theta) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} x_i x_i' f_i\left(\xi_i(\theta)\right),$$

sendo $f_i(\xi_i(\theta))$ a densidade condicional da variável resposta (y_i) avaliada no θ-ésimo quantil condicional. No caso i.i.d. as funções $f_i(\xi_i(\theta))$ são idênticas e o "Huber Sandwich" se iguala à expressão em Λ_θ.

2.2. Modelos

Nesta aplicação, para estimar a equação de rendimento, foram utilizados dois tipos de modelos: o primeiro é um modelo que pressupõe uma relação linear entre rendimento e escolaridade do tipo introduzido por Mincer (1974). Com este modelo, aqui denominado modelo tipo 1, estima-se o efeito de cada ano de escolaridade completo sobre o *logaritmo* do salário. Este efeito é quantificado pelo valor estimado do parâmetro da variável educação (ver nota de rodapé 1 para a definição da taxa de retorno em educação sobre o *salário*). Será também estimado um modelo mais geral, o qual possibilita a existência de não-linearidades na relação entre educação e salário [7]. Neste modelo a variável que quantifica os anos completos de estudo é transformada em diversas variáveis indicadoras que passam a distinguir os seguintes níveis ducacionais: fundamental menor, fundamental maior, médio e superior. De acordo com Lam e

[6] Segundo Koenker e Portnoy (op. cit.), o termo "função sparsity" foi criado por Tukey (1965), enquanto que Parzen (1979) denomina s(θ) de "função densidade-quantil".

[7] Ver Soares e Gonzaga (1999).

Schoeni (1993) essa especificação mais flexível é necessária devido à grande dispersão da escolaridade no Brasil. Ao adotar este modelo, aqui denominado por modelo tipo 2, considera-se que a taxa de retorno em educação não é constante para todo o ciclo escolar, ou seja, um ano de estudo completo para um determinado nível educacional repercutirá de forma diferente no salário relativamente a outro.

No modelo do tipo 1 o quantil condicional do logaritmo do rendimento-hora da mulher é dado por:

$$Q_\theta(y_i \mid Edu, X) = \mathrm{Edu}\beta_\theta + X\lambda_\theta \qquad (1)$$

onde "Edu" representa os anos completos de estudo e o vetor X as variáveis explanatórias, as quais incluem os principais atributos determinantes do salário e as características familiares das mulheres (estas variáveis serão descritas na próxima seção). Para o modelo do tipo 2, por sua vez, tem-se:

$$Q_\theta(y_i \mid Edu_1, Edu_2, Edu_3, Edu_4, X) = \sum_{j=1}^{4} \mathrm{Edu}_j \beta_{j\theta} + X\lambda_\theta \qquad (2)$$

onde o termo $\sum_{j=1}^{4} \mathrm{Edu}_j \lambda_{j\theta}$ indica o efeito das variáveis indicadoras educacionais no logaritmo do salário, tal como utilizado por Ueda e Hoffman(2002): Edu_0: analfabeta ou com menos de um ano de estudo (nível de referência); Edu_1: fundamental menor (1 a 4 anos de estudo); Edu_2: fundamental maior (5 a 8 anos de estudo); Edu_3: médio (9 a 11 anos de estudo) e Edu_4: superior (12 ou mais anos de estudo)[8].

No modelo tipo 1, o efeito marginal de uma determinada variável preditora em um quantil condicional de ordem θ da variável dependente pode ser obtido pela derivada parcial correspondente. Desta maneira, o efeito da educação no *salário,* denominado de taxa de retorno em educação, é definido como:

$$100 \times [\exp(\beta_\theta) - 1],$$

onde β_θ é o coeficiente da variável educação conforme especificado em (1).

[8] Esta divisão não implica que a pessoa tenha concluído um determinado nível educacional e sim que está cursando uma das séries do mesmo.

Para o cálculo da taxa de retorno em educação no modelo do tipo 2 seguimos o procedimento usado em Ueda e Hoffman (*op. cit.*). Neste artigo são calculadas as taxas *anuais* de retorno em educação para cada um dos quatro níveis educacionais (denominadas aqui por T_j, com $j = 1,2,3 \text{ e } 4$) conforme as fórmulas abaixo:

$$T_1 = 100 \times [\exp(\beta_1 / 2{,}5) - 1] \qquad \text{e} \tag{3}$$

$$T_j = 100 \times \left[\exp\left(\frac{\beta_j - \beta_{j-1}}{\mu_j - \mu_{j-1}} \right) - 1 \right] \quad \text{para } j = 2,3,4 \tag{4}$$

onde μ_j e β_j são a média dos anos de estudo em cada nível educacional e o coeficiente da variável indicadora em particular, respectivamente e, μ_{j-1} e β_{j-1}, suas primeiras defasagens. Para o primeiro nível educacional (fundamental menor: 1 a 4 anos de estudo) a média é 2,5, para o segundo nível educacional (fundamental maior: 5 a 8 anos de estudo) é 6,5 e assim por diante. A normalização do coeficiente nas fórmulas pela média de anos de estudo de cada nível educacional é necessária para se ter uma taxa anual. Como os níveis educacionais são cumulativos, tem-se igualmente de subtrair o correspondente efeito do nível educacional imediatamente inferior àquele para o qual se está calculando a taxa de retorno.

3. Dados e estatísticas descritivas

Os dados utilizados nesta aplicação são provenientes da Pesquisa Nacional por Amostra de Domicílio (PNAD) para os anos de 1992 e 1999. As subamostras aqui analisadas totalizam 24588 observações no ano de 1992 e 30749 em 1999 de mulheres com idade entre 20 e 64 anos que satisfazem às seguintes restrições: (a) trabalharam no ano da pesquisa; (b) tiveram rendimento mensal do trabalho positivo; (c) não eram empregadoras; e (d) tiveram todas as informações de interesse completas no banco de dados nas PNADs. As variáveis utilizadas englobam características pessoais e de emprego.

A Tabela 1 apresenta os valores observados das médias, desvios padrão e medianas, para as variáveis utilizadas na regressão, ponderados pelo fator de expansão amostral associados a cada observação. Em 1992 observa-se, por exemplo, que a média de anos de estudo foi de 5,12 anos e a média salarial R$1,79/hora, enquanto que em 1999 estes números são 6,04 anos e R$2,37/hora, respectivamente. Esses valores revelam o aumento médio da escolarização e da

remuneração do trabalho das mulheres no Brasil durante a década de 1990. Especialmente quanto à remuneração, os valores revelam ainda a grande assimetria positiva da distribuição salarial nos dois anos analisados, com a mediana bem menor do que a média e ainda um aumento da dispersão relativa, medida através do coeficiente de variação, no ano 1999, ou seja, a concentração de renda observada em 1992 persiste e mesmo se agrava em 1999.

A idade média das mulheres que trabalham aumentou de 38,53 anos em 1992 para 39,43 anos em 1999. Quanto à participação em sindicato, houve um pequeno aumento de 13% para 15% no período e o trabalho formal manteve sua participação em 37%.

Tabela 1 - Estatísticas básicas para as variáveis da amostra.
Mulheres de 20 a 64 anos, com renda do trabalho positiva (Brasil 1992 e 1999)

Variável	1992			1999		
	Média	Des.Padrão	Mediana	Média	Des.Padrão	Mediana
Salário/hora[1] (R$)	1,79	3,26	0,93	2,37	4,77	1,25
Ln_Salário[2]	0,30	1,05	0,25	0,56	0,96	0,48
Edu (Anos de estudo)	5,12	4,26	4,00	6,04	4,38	5,00
Edu_0*	0,19	-	-	0,14	-	-
Edu_1*	0,38	-	-	0,32	-	-
Edu_2*	0,22	-	-	0,26	-	-
Edu_3*	0,13	-	-	0,19	-	-
Edu_4*	0,08	-	-	0,09	-	-
Idade	38,53	11,69	37,00	39,43	11,70	38,00
Sindicato*	0,13	-	-	0,15	-	-
Trab. Formal*	0,37	-	-	0,37	-	-
Raça (branca)*	0,57	-	-	0,56	-	-
Região Norte*	0,04	-	-	0,05	-	-
Região Nordeste*	0,26	-	-	0,26	-	-
Região Sudeste*	0,46	-	-	0,45	-	-
Região Sul*	0,17	-	-	0,16	-	-
Região Centro-Oeste*	0,07	-	-	0,07	-	-
Zona Urbana*	0,81	-	-	0,82	-	-

***Variáveis binárias. Fonte dos dados: IBGE-PNADs 1992 e 1999. (1) Valor do salário/hora a preço de 1999, corrigido usando o INPC - Índice Nacional de Preço ao Consumidor. (2) Valor do logaritmo natural do salário hora.**

Resultados econométricos

O modelo de regressão quantílica permite estimar qualquer ponto da distribuição condicional da variável resposta dadas as variáveis explicativas. Neste estudo restringimos a estimação aos quantis {0,10; 0,25; 0,50; 0,75; 0,90}, os quais nos dão uma visão abrangente da relação entre o salário e seus determinantes. A variável dependente em todas as regressões é o logaritmo natural do salário/hora, definido como o salário semanal dividido pelo número de horas trabalhadas por semana e as equações estimadas estão ponderadas pelo fator de expansão

amostral da pesquisa. Tanto no modelo tipo 1 quanto no modelo tipo 2, o vetor X contém as seguintes variáveis explicativas: idade[9], idade ao quadrado e diversas variáveis indicadoras para as categorias das seguintes variáveis dicotômicas: a mulher trabalha no setor formal[10], participa de sindicato, raça (branca), reside em zona urbana e ainda as variáveis indicadoras regionais (sendo o sudeste a referência). Será enfatizada na análise o retorno em educação e particularmente o padrão de variação desse retorno ao longo dos diversos quantis da distribuição salarial. Os resultados serão apresentados a seguir para ambos os modelos considerados.

Modelo tipo 1

Os resultados de regressão quantílica e de MQO para os anos de 1992 e 1999 estão apresentados nas Tabelas 2 e 3. Todos os coeficientes são estatisticamente significativos ao nível de 1% e seus sinais correspondem aos esperados pela teoria econômica. Os coeficientes de idade e idade ao quadrado sendo positivo e negativo, respectivamente, em todos os quantis, estão de acordo com a teoria do capital humano, que estabelece que os rendimentos seguem uma curva parabólica, devido à depreciação do capital humano do trabalhador que, com o envelhecimento, tem o seu desempenho diminuído. Mulheres que participam de sindicato, da raça branca e residem em área urbana recebem, em geral, salários mais altos. As estimativas negativas das variáveis indicadoras regionais indicam salários mais altos no sudeste relativamente às demais regiões. Nota-se ainda que o diferencial é maior para aquelas residentes na região Nordeste, ou seja, esta região apresenta os mais baixos salários. O efeito de exercer atividades no setor formal vai decrescendo ao longo dos quantis, tanto em 1992 quanto em 1999. Porém, há uma redução do impacto desta variável de 1992 para 1999.

Com relação ao efeito da educação no salário, nosso foco de análise, observa-se um crescimento gradativo do coeficiente da variável "Edu" ao longo dos quantis, tanto em 1992 quanto em 1999. Sendo que em 1999 o crescimento foi um pouco superior ao de 1992 (Tabelas 2 e 3). Os resultados demonstram que a estimativa do coeficiente da variável educação na média condicional (por MQO) difere significativamente daquelas encontradas para os diversos quantis

---

[9] A *proxy* para experiência comumente usada é calculada por min {idade - educação - 6, idade -18} (ver Buchinsky, 1996). No entanto, para o caso brasileiro esta *proxy* não é recomendável, dado haver uma grande incidência de repetências escolares e também o fato de o ingresso ao mercado de trabalho ser muitas vezes tardio devido ao desemprego. No caso das mulheres, em particular, também há a saída temporária do mercado de trabalho para a criação dos filhos. Para o Brasil, a variável idade é, portanto, mais apropriada como *proxy* para experiência. Esta foi a especificação usada em Lam e Schoeni (*op. cit.*) e Ueda e Hoffmann (*op. cit.*).

[10] Estão incluídas nesta categoria todas as mulheres que possuíam carteira assinada, eram militares ou trabalhavam no serviço público na semana de referência da pesquisa.

(usando a regressão quantílica), em especial nas caudas da distribuição (quantis inferiores e superiores). Nota-se ainda que o diferencial entre as estimativas dos coeficientes da variável educação ao longo dos quantis é maior no ano de 1999 relativamente a 1992.

Tabela 2 - Estimativas dos parâmetros no modelo (tipo 1) de regressão quantílica e de MQO, para o ano 1992

Variáveis	Quantis					MQO
	0,10	0,25	0,50	0,75	0,90	
Const	-2,4838	-2,0364	-1,7234	-1,4001	-1,0009	-1,7956
Idade	0,0404	0,0421	0,0476	0,0514	0,0500	0,04924
Idade^2	-0,0004	-0,0004	-0,0005	-0,0005	-0,0004	-0,0005
Edu	0,0916	0,1001	0,1104	0,1199	0,1271	0,1124
Trab. Formal	0,5836	0,3749	0,2191	0,0933	-0,0366	0,2398
Sindicato	0,2082	0,2300	0,2387	0,2758	0,2942	0,2543
Raça (branca)	0,1647	0,1113	0,1189	0,1416	0,1520	0,1391
Norte	-0,2581	-0,2224	-0,1681	-0,1169	-0,1184	-0,1824
Nordeste	-0,7188	-0,5479	-0,4193	-0,3406	-0,3060	-0,4674
Sul	-0,0937	-0,0810	-0,1009	-0,1139	-0,1132	-0,1007
Centro-Oeste	-0,0828	-0,0770	-0,0866	-0,1186	-0,0654	-0,0864
Zona Urbana	0,2600	0,2410	0,1917	0,1898	0,2250	0,2159

Nota: Todos os coeficientes são estatisticamente significativos ao nível de 1%.

Tabela 3 - Estimativas dos parâmetros no modelo (tipo 1) de regressão Quantílica e de MQO, para o ano 1999

Variáveis	Quantis					MQO
	0,10	0,25	0,50	0,75	0,90	
Const	-2,0655	-1,9138	-1,6478	-1,4049	-1,1520	-1,6592
Idade	0,0392	0,0452	0,0501	0,0576	0,0648	0,0516
Idade^2	-0,0004	-0,0004	-0,0004	-0,0005	-0,0006	-0,0005
Edu	0,0841	0,0975	0,1122	0,1230	0,1324	0,1114
Trab. Formal	0,4062	0,2802	0,1401	-0,0071	-0,1636	0,1394
Sindicato	0,2213	0,2270	0,2465	0,2336	0,2431	0,2402
Raça (branca)	0,1169	0,1295	0,1325	0,1549	0,1636	0,1456
Norte	-0,2123	-0,1945	-0,1938	-0,1819	-0,1264	-0,2005
Nordeste	-0,4555	-0,4018	-0,3734	-0,3761	-0,3857	-0,4085
Sul	-0,1337	-0,1474	-0,1700	-0,1909	-0,1963	-0,1758
Centro-Oeste	-0,1351	-0,1375	-0,1517	-0,1371	-0,0991	-0,1335
Zona Urbana	0,1808	0,1673	0,1302	0,1393	0,1636	0,1626

Nota: Todos os coeficientes são estatisticamente significativos ao nível de 1%.

Modelo tipo 2

Os resultados de regressão quantílica e de MQO para os anos de 1992 e 1999 são apresentados nas Tabelas 4 e 5. Tal como no modelo tipo 1, todos os coeficientes são estatisticamente significativos ao nível de 1% e possuem os sinais esperados. Temos igualmente

que os resultados para os quantis são distintos daquele para a média condicional. No modelo tipo 2, os coeficientes estimados das quatro variáveis indicadoras educacionais, tal como era de se esperar, estão ordenados, em cada quantil, ou seja, o coeficiente de menor valor é o referente ao menor grau de escolarização e o maior coeficiente é o daquela variável indicadora que representa o maior nível de escolarização. Para as demais variáveis os coeficientes apresentam as mesmas tendências das estimativas obtidas usando o modelo tipo 1.

Tabela 4 - Estimativas dos parâmetros no modelo (tipo 2) de regressão Quantílica e de MQO, para o ano 1992

Variáveis	Quantis					MQO
	0,10	0,25	0,50	0,75	0,90	
Const	-2,4930	-1,8866	-1,5362	-1,1627	-0,8256	-1,6712
Idade	0,0395	0,0406	0,0440	0,0459	0,0465	0,0473
Idade^2	-0,0004	-0,0004	-0,0004	-0,0004	-0,0004	-0,0005
Edu_1	0,3529	0,2319	0,2308	0,2670	0,2783	0,2738
Edu_2	0,5767	0,4566	0,5072	0,5828	0,6096	0,5682
Edu_3	0,9018	0,8797	0,9751	1,1098	1,1617	1,0302
Edu_4	1,3741	1,4426	1,5852	1,7315	1,8384	1,6070
Trab. Formal	0,6219	0,4077	0,2381	0,0988	-0,0228	0,2609
Sindicato	0,2057	0,2132	0,2368	0,2751	0,2713	0,2485
Raça (branca)	0,1652	0,1208	0,2381	0,1492	0,1887	0,1547
Norte	-0,2221	-0,2200	-0,1543	-0,1325	-0,1236	-0,1795
Nordeste	-0,7139	-0,5662	-0,4402	-0,3848	-0,3250	-0,4857
Sul	-0,0842	-0,0805	-0,0969	-0,1243	-0,1077	-0,1054
Centro-Oeste	-0,0683	-0,0768	-0,0918	-0,1412	-0,0808	-0,0835
Zona Urbana	0,2666	0,2514	0,2265	0,2183	0,2530	0,2430

Nota: Todos os coeficientes são estatísticamente significativos ao nível de 1%.

Tabela 5- Estimativas dos parâmetros no modelo(tipo 2) de regressão Quantílica e de MQO, para o ano 1999

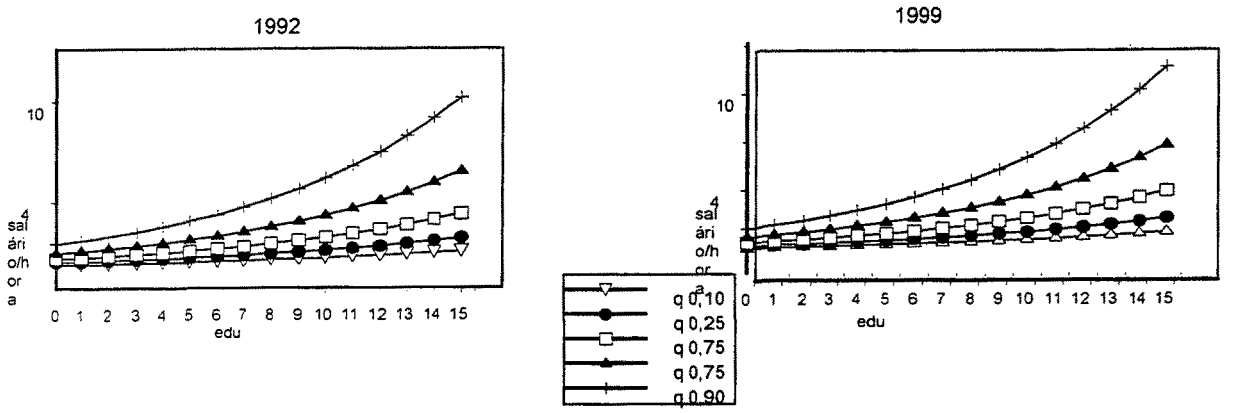
Variáveis	Quantis					MQO
	0,10	0,25	0,50	0,75	0,90	
Const	-1,8957	-1,5704	-1,2844	-1,0780	-0,7402	-1,3623
Idade	0,0394	0,0406	0,0463	0,0548	0,0604	0,0496
Idade^2	-0,0004	-0,0004	-0,0004	-0,0005	-0,0006	-0,0005
Edu_1	0,1478	0,1381	0,1299	0,1704	0,1470	0,1592
Edu_2	0,3511	0,3327	0,3483	0,4083	0,3977	0,3903
Edu_3	0,6574	0,7183	0,8301	0,9841	1,0020	0,8551
Edu_4	1,3243	1,4785	1,6113	1,7780	1,8541	1,6207
Trab. Formal	0,4230	0,3001	0,1560	0,0018	-0,1531	0,1539
Sindicato	0,1938	0,1872	0,1955	0,1862	0,2337	0,1961
Raça (branca)	0,1087	0,1235	0,1402	0,1505	0,1691	0,1484
Norte	-0,1888	-0,2104	-0,2027	-0,1937	-0,1237	-0,1960
Nordeste	-0,4738	-0,4258	-0,3966	-0,4122	-0,4146	-0,4333
Sul	-0,1243	-0,1365	-0,1528	-0,1756	-0,1788	-0,1603
Centro-Oeste	-0,1500	-0,1450	-0,1488	-0,1440	-0,1043	-0,1360
Zona Urbana	0,2135	0,2041	0,1737	0,1981	0,1902	0,2012

Nota: Todos os coeficientes são estatísticamente significativos ao nível de 1%.

O Gráfico 1[11] mostra os salários ajustados para os quantis {0,10; 0,25; 0,50; 0,75; 0,90} em função de anos de estudo, mantendo as demais variáveis nos seus valores médios, para os anos de 1992 e 1999.

Os resultados mais aparentes neste gráfico são os distintos retornos em educação (inclinação das diversas curvas ao longo dos quantis), bem como a existência de uma maior dispersão do salário esperado para mulheres com maior escolaridade. Ou seja, pessoas de baixa escolaridade tendem a receber salários mais baixos e mais homogêneos, enquanto que para pessoas de alta escolaridade há uma maior heterogeneidade dos níveis salariais.

Gráfico 1 - Salário ajustado em função de anos de estudo, para mulheres entre 20 e 64 anos, por quantil. Brasil 1992 - 1999



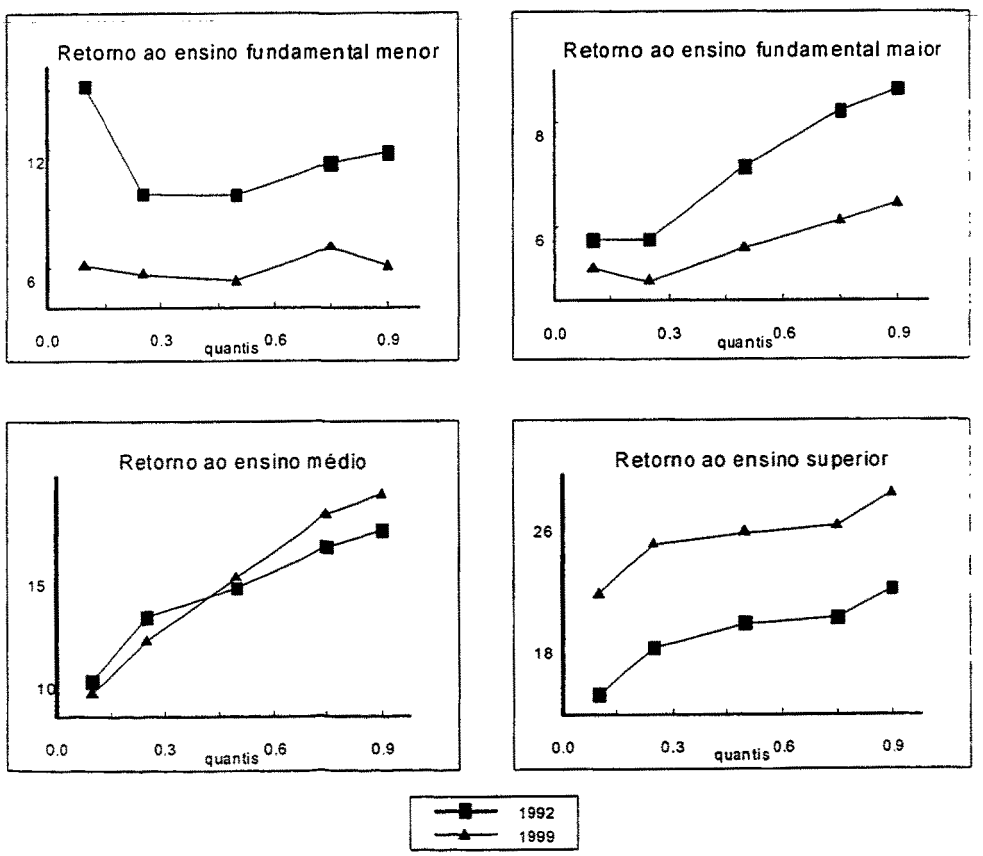
Outra conclusão evidente é o crescimento do salário em todos os quantis conforme o nível educacional se torna mais elevado. No entanto, há algumas diferenças neste crescimento entre os dois anos. Nota-se que para o ano de 1999 a curva do salário esperado para o quantil 0,90 é mais inclinada que a respectiva curva para o ano de 1992, para níveis educacionais mais elevados. Isto indica um crescimento no retorno em educação no período para as mulheres com maior renda salarial e nível educacional mais elevado. Isto também se verifica em menor proporção para o quantil 0,75. Percebe-se ainda neste gráfico que nos demais pontos da distribuição, em especial os quantis abaixo da mediana, houve pouca variação no crescimento do salário em função dos anos de estudo.

No Gráfico 2 encontra-se ilustrado o retorno em educação para os quatro níveis educacionais nos anos de 1992 e 1999 (computado pelas fórmulas dadas nas equações (3) e

[11] Ambos os gráficos (1 e 2) foram construídos com base nos resultados do modelo tipo 2, cuja especificação é menos restritiva.

(4)). Observa-se que em 1999 o retorno em educação tende a aumentar à medida que se move para quantis mais elevados da distribuição salarial para a maioria dos níveis educacionais, com exceção apenas do primeiro (fundamental menor), para o qual há uma pequena variação da taxa de retorno ao longo dos quantis. Em 1992 por sua vez, para o fundamental menor, ocorreu uma queda acentuada no retorno em educação no quantil 0,25. A partir da mediana nota-se uma reversão desta tendência, havendo um crescimento do retorno em educação à medida que se eleva o número de anos de estudo, semelhante ao encontrado para 1999. Nota-se também que nos dois anos estudados há um comportamento diferente do retorno em educação para as mulheres com mais baixa escolaridade, visto que tanto no primeiro nível como no segundo nível de escolaridade (*fundamental menor e maior*) o retorno diminui do quantil 0,10 para o quantil 0,25 e, em alguns casos, até a mediana, passando então a aumentar até o quantil 0,90, como pode ser visualizado no Gráfico 2 já referido. Nestes dois níveis de escolaridade temos ainda que os retornos, em todos os quantis, são maiores em 1992 do que em 1999. Isto indica que no período houve uma queda no retorno em educação para aqueles indivíduos menos qualificados.

Gráfico 2 - Taxa de retorno em educação para mulheres entre 20 e 64 anos, em vários níveis de educação, por quantil. Brasil - 1992 - 1999.



Para o terceiro nível de educação, em ambos os anos, os resultados demonstram que o retorno aumenta à medida que nos movemos para a cauda direita da distribuição salarial, sendo que até o quantil 0,25 os retornos em 1992 são maiores do que em 1999. Após a mediana ocorre uma reversão, ou seja, os retornos passam a ser maiores em 1999. Para o nível educacional mais elevado tem-se que os retornos são maiores em 1999 comparados a 1992, em todos os quantis. De forma geral, os resultados no Gráfico 2 apontam para uma valorização de profissionais mais qualificados ao longo do período, ou seja, pessoas mais qualificadas passaram a ganhar mais e pessoas menos qualificadas passaram a ganhar menos. Analisando ainda os retornos em termos relativos, pode-se ainda destacar o fato de que os retornos são maiores nos níveis de educação mais elevados quando comparado com os retornos nos mais baixos níveis de escolaridade.

5. Conclusões

Este trabalho analisou as mudanças na distribuição salarial das mulheres no Brasil entre os anos de 1992 e 1999, em particular o retorno em educação, que vem sofrendo grandes transformações ao longo das últimas décadas. O método de regressão quantílica possibilitou uma melhor caracterização deste retorno em vários pontos da distribuição salarial das mulheres, provendo uma visão mais global da relação entre educação e rendimentos das mesmas. As principais conclusões são sumarizadas a seguir:

- O efeito da educação não é o mesmo ao longo da distribuição condicional do salário para ambos os anos estudados. Desta forma, a média condicional nos fornece apenas uma visão limitada da realidade.
- Apesar de algumas semelhanças no comportamento do retorno em educação em ambos os anos estudados, o mesmo é maior no ano de 1999 relativamente ao ano de 1992 para os quantis mais elevados da distribuição salarial;
- Para os mais baixos níveis de escolaridade, tanto em 1992 quanto em 1999, o retorno em educação diminuiu do quantil 0,10 para o quantil 0,25.
- O retorno em educação para as mulheres que possuem no máximo oito anos de escolaridade (ensino fundamental menor e maior) teve um decréscimo de 1992 para 1999, indicando que em 1999 a qualificação passou a ser mais valorizada.
- Para as mulheres que conseguiram fazer pelo menos um ano do curso superior, o retorno em educação é sempre superior em 1999, indicando a tendência de crescimento do retorno em

educação para os quantis mais elevados e para aquelas que investem em maior qualificação.

Os resultados desta aplicação indicam, portanto, que a variável educação afeta de forma distinta a distribuição condicional do salário das mulheres, sendo o retorno em educação diferente para os quantis estimados. A análise também revelou diferentes tendências do retorno em educação entre os anos estudados (1992 e 1999), em particular houve uma valorização do trabalho de profissionais mais qualificados (com maior número de anos completos de estudo) no período recente.

Referências bibliográficas

ARAÚJO, T. P. (1996), Returns to education in segmented labour markets. *Textos para discussão nº 367*. PIMES. UFPE. Pernambuco.

BARROS, R., JATOBÁ, J. e MENDONÇA, R. (1997), A evolução da participação de mulheres no mercado de trabalho: uma análise de decomposição. In: Anais do IV Encontro Nacional de Estudos do Trabalho, São Paulo, Setembro de 1995. Associação Brasileira de Estudos do Trabalho. Rio de Janeiro.

BECKER, G.(1964), Human Capital: A theorical and empirical analysis, with special reference to education. Nova York. Columbia University Press.

BUCHINSKY, M.(1994), Changes in the U.S. wage structure 1963-1987: Application of quantile regression. Econometrica, 62, 405-458.

BUCHINSKY, M. (1996), Women's return to education in the U.S.: Exploration by quantile regression with nonparametric sample selection correction. Yale University and NBER.

BUCHINSKY, M. (1997), Recent advances in quantile regression: A pratical guideline for for empirical research. Brown University and NBER.

HARTOG, J., PEREIRA, P. T. e VIEIRA, J. A. C. (1999), Changing returns to education in Portugal during the 1980s and early 1990s: OLS and quantile regression estimators. Tinbergen Institute Discussion Papers, No 2.

JACOBSEN, J. P. (1999), Labor force participation. In: Women at the end of the millennium: What we now, what we need to know. The Quartely Review of Economics and Finance, 39, 597-610.

JATOBÁ, J. (1993), Brazilian women in the metropolitan labor force: A time series study across region and household status. Série Seminários n0 27/93. IPEA. Rio de Janeiro.

KOENKER, R. e BASSET, G. (1978), Regression Quantiles. Econometrica, 46, 33-50.

KOENKER, R. e BASSET, G. (1982), Robusts tests for heteroscedasticity based on regression quantiles. Econometrica, 50, 43-61.

KOENKER, R. e HALLOCK,K.F. (2001), Quantile regression: An introduction.
www.econ.uiuc.edu/~roger/research/intro/intro.html.

KOENKER, R. e MACHADO J. (1999), Goodness of fit and related inference process for quantile regression. Journal of American Statistical Association, 94, 1296-1310.

KOENKER, R.; PORTNOY, S. (1997). Quantile Regression. 5ª Escola de Modelos de Regressão. Associação Brasileira de Estatística.

LAM, D. e SCHOENI R. (1993), Effects of family background on earnings and returns to schooling: evidence from Brazil. Journal of Political Economy, 101, 710-740.

LEAL, C. I. S. e VERLANG, S. R. C. (1991), Retornos em Educação no Brasil: 1976/89. Pesquisa e Planejamento Econômico, 21, 559-574.

MINCER, J. (1974), Schooling, Experience and Earnings. New York: National Bureau of Economic Research.

MOSTELLER, F. e TUKEY, J. (1977), Data Analysis and Regression: A second Course in Statistics. Addison-Wesley, Reading Mass.

PARZEN, E. (1979), Nonparametric statistical data modeling. Journal of American Statistical Association, 74, 105-123.

SANTOS, E. C. e SEDLACECK, G. L. (1990), A evolução da participação feminina no mercado de trabalho. Revista de Econometria, 10, 225-241.

SOARES, R. R. e GONZAGA, G. (1999), Determinação de salários no Brasil: Dualidade ou não-linearidade no retorno à educação? Revista de Econometria, 19(2).

STELCNER, M., SMITH, J. B., BRESLAW, J. A. e MONETTE, G. (1992), Labour force behavior and earnings of Brazilian women and men, 1980, in Case Studies on Women's Employment and Pay in Latin America (ed.) Psacharopoulos, G. e Tzannatos, Z. (Washington: World Bank), 39-88.

TUKEY, J. (1965), Which part of the sample contains the information? Proccedings of the National Academy, 53, 127-134.

UEDA, E. M. e HOFFMANN, R. (2002), Estimando o retorno da educação no Brasil. Economia Aplicada, 2, 209-238.

ZYLBERSTAJN, H., PAGOTO, C. S. e PASTORE, J. (1985), A mulher e o menor na força de trabalho. Nobel: Ministério do Trabalho, Brasil.

Abstract

Schooling is a major factor in determining the wage structure of workers and the main task of this article is to study the relation between education and women salaries (response variable) in Brazil in the 90s. The analysis will be based on the quantile regression technique. The results point to different solutions for the quantiles, thus indicating different responses of the dependent variable to the education level at distinct points of its conditional distribution. The results also reveal that the return to education is higher at the right tail of the conditional distribution of the response variable and has increased over the period studied. It is also noteworthy the fact that skilled females experienced an increase in the return to education while the opposite happened for less qualified ones.

Agradecimentos

Ana Katarina T. de N. Campêlo menciona e agradece o suporte financeiro através de Bolsa de Produtividade em Pesquisa do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Análise de resultados da prova tuberculínica: uso de métodos Bayesianos

Edson Zangiacomi Martinez*
Jorge Alberto Achcar*
Antonio Ruffino-Netto*

Resumo

A distribuição dos diâmetros das reações tuberculínicas usualmente é dada pela mistura de duas ou mais distribuições de probabilidade. Considerando dados de provas tuberculínicas realizadas em Ribeirão Preto, propomos neste artigo um modelo Bayesiano que pressupõe que os diâmetros das reações se comportam conforme uma mistura de duas distribuições normais. Quando comparado com alternativas da literatura, o modelo aqui proposto tem a vantagem de poder ser facilmente implementado e a sua capacidade de introduzir a informação de covariáveis. Usamos métodos de simulação baseados em algoritmos Markov Chain Monte Carlo - MCMC para obter as inferências *a posteriori* de interesse. Com o modelo ajustado, pudemos descrever como os diâmetros das reações tuberculínicas distribuem-se de acordo com as diferentes faixas etárias.

Palavras chave: tuberculose, prova tuberculínica, métodos Bayesianos.

* Endereço para correspondência: Departamento de Medicina Social, FMRP – USP, Faculdade de Medicina de Ribeirão Preto, Av. Bandeirantes, 3900, 14049-900, Ribeirão Preto, SP.

1. Introdução

A tuberculose é um dos mais graves problemas de saúde pública atualmente no mundo (WHO, 2003). A prova tuberculínica é importante instrumento de diagnóstico, sendo sua leitura feita através da medida do diâmetro da enduração. Se houver boa correlação entre eritema e enduração, tal informação poderá ser muito útil como meio auxiliar para o diagnóstico da doença. Carneiro (1962) assinala que no Japão, na sua época, a leitura da prova era feita através do eritema, e não da enduração.

Quando se analisa a distribuição dos diâmetros das reações tuberculínicas obtidas em uma população e dispostas na forma de histogramas de freqüências relativas, é possível obter diversas informações. Uma dessas informações é relativa ao comportamento da população face a infecções por micobactérias. Os padrões observados variam de região para região em decorrência da maior ou menor prevalência de reações fracas. Em geral, esse comportamento é dado pela mistura de duas ou mais distribuições de probabilidade. A experiência acumulada durante muitos anos permite atribuir essas reações fracas à ocorrência de reações cruzadas devidas a infecções por micobactérias atípicas.

Algumas populações, especialmente em países de regiões temperadas e subtropicais, são "puras", isto é, infectadas apenas pelo *Mycobacterium tuberculosis*. Em regiões tropicais é comum a existência de populações "impuras", isto é, infectadas pelo *Mycobacterium tuberculosis* e outras micobactérias, onde o poder discriminatório do teste tuberculínico é menor.

No presente artigo, consideramos os resultados de 35 680 testes tuberculínicos realizados em Ribeirão Preto, Estado de São Paulo (dados introduzidos por Ruffino-Netto et al., 1977). O teste foi aplicado, segundo a técnica preconizada pela Divisão Nacional de Tuberculose em conformidade com a Organização Mundial de Saúde, por operadores padronizados. A tuberculina usada foi o derivado protéico purificado conhecido como PPD rt23, lote 23, na dose de duas unidades de tuberculina. De acordo com a enduração, os resultados do teste são classificados como não-reatores (diâmetros relativamente pequenos) e reatores. A população de estudo foi constituída pelas pessoas de qualquer idade, residentes no município de Ribeirão Preto, que se matricularam em um centro de saúde da cidade para qualquer fim, durante 12 meses consecutivos.

Os histogramas construídos a partir destes dados e apresentados na Figura 1, sugerem uma população "impura", ou seja, há uma evidência de multimodalidade na distribuição dos diâmetros das provas tuberculínicas. Esse fato é supostamente devido à ocorrência de reações cruzadas

causadas por infecções por micobactérias atípicas. Desta forma, a modelagem matemática destes dados deve prever uma mistura de distribuições. A estimação dos parâmetros de uma mistura de duas distribuições normais por métodos clássicos foi estudada por diversos autores, como Cohen (1967), Day (1969) e Dick e Bowden (1973). O uso de métodos estatísticos Bayesianos para modelos de misturas tem sido considerado por muitos estatísticos como, por exemplo, Neal (1991), Diebolt e Robert (1994), Escobar e West (1995), Richardson e Green (1997) e Roberts et al. (1998).

No presente trabalho, objetivamos apresentar uma modelagem Bayesiana para a análise da distribuição dos diâmetros das reações tuberculínicas, pressupondo que estes dados introduzidos por Ruffino-Netto et al. (1977) seguem uma mistura de duas distribuições normais. Quando comparado com alternativas da literatura, o modelo aqui proposto tem a vantagem de poder ser facilmente implementado e a capacidade de introduzir a informação de uma covariável.

2. O modelo estatístico

No presente modelo, vamos considerar a idade de cada indivíduo como covariável. Obedecendo à estrutura original do banco de dados introduzido por Ruffino-Neto et al. (1977), cada indivíduo pertence a uma das dez faixas etárias exibidas na Tabela 1. Seja X uma variável aleatória associada aos diâmetros de enduração da prova tuberculínica, e seja x uma observação desta variável, de forma que x_{il} denota o resultado do teste para o *l*-ésimo indivíduo da *i*-ésima faixa etária, i = 1,2,...,10 e $l = 1,2,\ldots,n_i$. Assim, associamos a x_{il} uma mistura de duas distribuições normais (ver, por exemplo, Titterington et al., 1985) com função densidade de probabilidade dada por

$$f(x_{il}) = \sum_{j=1}^{2} p_{ji}\phi_j\left(x_{il} \mid \mu_{ji}, \sigma_j^2\right) \qquad (1)$$

onde i = 1,2,...,10 (número de faixas etárias), $l = 1,2,\ldots,n_i$ (número de pessoas por faixa etária), $\phi_j\left(x_{il} \mid \mu_{ji}, \sigma_j^2\right)$ denota a função densidade de probabilidade normal $N\left(\mu_{ji}, \sigma_j^2\right)$ com média μ_{ji} e variância σ_j^2 dada por

$$\phi_j\left(x_{il} \mid \mu_{ji}, \sigma_j^2\right) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{1}{2\sigma_j^2}\left(x_{il} - \mu_{ji}\right)^2\right] \tag{2}$$

para $j = 1,2$ e p_{ji} é proporção do componente j na mistura (as probabilidades de incidência), sendo que $p_{1i} + p_{2i} = 1$. Assumindo que a covariável W (faixa etária) influencia as médias μ_{ji} e as probabilidades de incidência $p_{ji}, j = 1,2$, escrevemos as relações

$$\mu_{ji} = \alpha_j + \beta_j w_i \tag{3}$$

e

$$p_{1i} = \frac{e^{\gamma + \tau w_i}}{1 + e^{\gamma + \tau w_i}} \tag{4}$$

(uma ligação logito), sendo $p_{2i} = 1 - p_{1i}, i = 1,\ldots,10$.

A função de verossimilhança para $\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \gamma, \tau)$ é dada por

$$L(\theta) = \prod_{i=1}^{10} \prod_{l=1}^{n_i} \left[\sum_{j=1}^{2} p_{ji} \phi_j\left(x_{il} \mid \mu_{ji}, \sigma_j^2\right)\right] \tag{5}$$

onde μ_{ji} e p_{ji} são dados em (3) e (4), respectivamente.

Para a obtenção de estimadores para os parâmetros, podemos usar métodos clássicos de inferência ou métodos Bayesianos. A estimação dos parâmetros pelo método da máxima verossimilhança torna-se complexa devido à soma observada na expressão (5) e aspectos de não identificabilidade. Por sua vez, a estimação Bayesiana é um método atraente de estimação, com a eliminação da soma observada em (5) pela introdução de variáveis latentes (Tanner e Wong, 1987) como será descrito a seguir.

3. Análise bayesiana do modelo

Para uma análise Bayesiana do modelo, vamos considerar as seguintes distribuições *a priori* para $\alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \gamma$ e τ :

$$\begin{aligned} &\alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2 \text{ : localmente uniformes;} \\ &\gamma \sim N(a_1, b_1^2),\ a_1 \text{ e } b_1 \text{ hiperparâmetros conhecidos;} \\ &\tau \sim N(a_2, b_2^2),\ a_2 \text{ e } b_2 \text{ hiperparâmetros conhecidos.} \end{aligned} \tag{6}$$

Sob o pressuposto de que estes parâmetros são independentes, temos que

$$\pi(\theta) \propto \exp\left(-\frac{(\gamma - a_1)^2}{2b_1^2}\right) \exp\left(-\frac{(\tau - a_2)^2}{2b_2^2}\right) \tag{7}$$

Usando a fórmula de Bayes (ver por exemplo, Box e Tiao, 1973), encontramos a distribuição *a posteriori* conjunta para θ :

$$\pi(\theta \mid X) \propto \pi(\theta) \prod_{i=1}^{10} \prod_{l=1}^{n_i} \sum_{j=1}^{2} p_{ji} \phi_j\left(x_{il} \mid \mu_{ji}, \sigma_j^2\right) \tag{8}$$

Para encontrarmos amostras da distribuição *a posteriori* conjunta (8), utilizamos amostradores de Gibbs (Gelfand e Smith, 1990). Para simplificação do algoritmo, introduzimos variáveis artificiais (ver, por exemplo, Tanner e Wong, 1987) para a eliminação da soma em (8). Definimos então $\mathbf{Z}_{il}^{T} = (\mathbf{Z}_{il1}, \mathbf{Z}_{il2})$, onde $\mathbf{Z}_{il1} + \mathbf{Z}_{il2} = 1, \quad i = 1, \ldots, 10, i = 1, \ldots, n_i$ e $\mathbf{Z}_{il1}$ tem distribuição de Bernoulli com probabilidade de sucesso h_{il1} dada por

$$h_{il1} = \frac{p_{1i} \phi_1\left(x_{il} \mid \mu_{1i}, \sigma_1^2\right)}{\sum_{j=1}^{2} p_{ji} \phi_j\left(x_{il} \mid \mu_{ji}, \sigma_j^2\right)} \tag{9}$$

Assim,

$$\pi(z_{il} \mid \theta, x_{il}) \propto h_{il1}^{z_{il1}} (1-h_{il1})^{z_{il2}} \tag{10}$$

onde $Z_{il2} = 1 - Z_{il1}$. De (9) e (10), temos

$$\pi(Z \mid \theta, X) \propto \frac{\prod_{i=1}^{10} \prod_{l=1}^{n_i} \prod_{j=1}^{2} \left[p_{ji} \phi_j \left(x_{il} \mid \mu_{ji}, \sigma_j^2 \right) \right]^{z_{ilj}}}{\prod_{i=1}^{10} \prod_{l=1}^{n_i} \sum_{j=1}^{2} p_{ji} \phi_j \left(x_{il} \mid \mu_{ji}, \sigma_j^2 \right)} \tag{11}$$

onde $Z' = (Z_{11}', Z_{12}', \ldots, Z_{10,n10}')$. Combinando (8) com (11) temos

$$\pi(Z \mid \theta, X) \propto \pi(\theta) \prod_{i=1}^{10} \prod_{l=1}^{n_i} \prod_{j=1}^{2} \left[p_{ji} \phi_j \left(x_{il} \mid \mu_{ji}, \sigma_j^2 \right) \right]^{z_{ilj}} \tag{12}$$

onde $\pi(\theta)$ é a distribuição *a priori* conjunta para $\theta' = (\alpha_1, \alpha_2, \beta_1, \beta_2, \sigma_1^2, \sigma_2^2, \gamma, \tau)$, $p_{1i} = \exp(\gamma + \tau W_i) / (1 + \exp(\gamma + \tau W_i))$, $p_{2i} = 1 - p_{1i}$, e $\mu_{ji} = \alpha_j + \beta_j w_i$, $j=1,2$, $i=1,\ldots,10$. Substituindo (7) em (12) e desenvolvendo o produto $\prod_{i=1}^{10} \prod_{l=1}^{n_i} \prod_{j=1}^{2} \left[p_{ji} \phi_j \left(x_{il} \mid \mu_{ji}, \sigma_j^2 \right) \right]^{z_{ilj}}$, temos

$$\begin{aligned} \pi(\theta \mid X, Z) \propto & \exp\left(-\frac{(\gamma - a_1)^2}{2b_1^2} \right) \exp\left(-\frac{(\tau - a_2)^2}{2b_2^2} \right) \left(\frac{1}{\sigma_1} \right)^r \left(\frac{1}{\sigma_2} \right)^{n_\bullet - r} \times \\ & \times \exp\left(-\frac{1}{2\sigma_1^2} \sum_{i=1}^{10} \sum_{l=1}^{n_i} z_{il1} (x_{il} - \mu_{1i})^2 - \frac{1}{2\sigma_2^2} \sum_{i=1}^{10} \sum_{l=1}^{n_i} z_{il2} (x_{il} - \mu_{2i})^2 \right) \times \\ & \times \frac{\exp(\gamma r + \tau s)}{\prod_{i=1}^{10} \prod_{l=1}^{n_i} \left(1 + e^{r + \tau w_i} \right)}, \end{aligned} \tag{13}$$

onde $n_\bullet = \sum_{i=1}^{10} n_i$, $r = \sum_{i=1}^{10}\sum_{l=1}^{n_i} z_{il1}$, $n_\bullet - r = \sum_{i=1}^{10}\sum_{l=1}^{n_i} z_{il2}$ e $s = \sum_{i=1}^{10}\sum_{l=1}^{n_i} z_{il1} w_i$.

A partir de (13), encontramos as distribuições condicionais necessárias para o amostrador de Gibbs:

$$\sigma_1^2 \mid \theta_{(\sigma_1^2)}, X, Z \sim IG\left[\frac{r}{2} - 1; \frac{1}{2}\sum_{i=1}^{10}\sum_{l=1}^{n_i} z_{il1}(x_{il} - \mu_{1i})^2\right] \tag{14}$$

$$\sigma_2^2 \mid \theta_{(\sigma_2^2)}, X, Z \sim IG\left[\frac{n_\bullet - r}{2} - 1; \frac{1}{2}\sum_{i=1}^{10}\sum_{l=1}^{n_i} z_{il2}(x_{il} - \mu_{2i})^2\right] \tag{15}$$

$$\alpha_1 \mid \theta_{(\alpha_1)}, X, Z \sim N\left[\frac{1}{r}\sum_{i=1}^{10}\sum_{l=1}^{n_i} z_{il1}(x_{il} - \beta_1 w_i); \frac{\sigma_1^2}{r}\right] \tag{16}$$

$$\alpha_2 \mid \theta_{(\alpha_2)}, X, Z \sim N\left[\frac{1}{n_\bullet - r}\sum_{i=1}^{10}\sum_{l=1}^{n_i} z_{il2}(x_{il} - \beta_2 w_i); \frac{\sigma_2^2}{n_\bullet - r}\right] \tag{17}$$

$$\beta_1 \mid \theta_{(\beta_1)}, X, Z \sim N\left[\frac{\sum_{i=1}^{10}\sum_{l=1}^{n_i} z_{il1} w_i (x_{il} - \alpha_1)}{\sum_{i=1}^{10}\sum_{l=1}^{n_i} z_{il1} w_i^2}; \frac{\sigma_1^2}{\sum_{i=1}^{10}\sum_{l=1}^{n_i} z_{il1} w_i^2}\right] \tag{18}$$

$$\beta_2 \mid \theta_{(\beta_2)}, X, Z \sim N\left[\frac{\sum_{i=1}^{10}\sum_{l=1}^{n_i} z_{il2} w_i (x_{il} - \alpha_2)}{\sum_{i=1}^{10}\sum_{l=1}^{n_i} z_{il2} w_i^2}; \frac{\sigma_2^2}{\sum_{i=1}^{10}\sum_{l=1}^{n_i} z_{il2} w_i^2}\right] \tag{19}$$

$$\pi(\gamma \mid \theta_{(\gamma)}, X, Z) \propto \exp\left(-\frac{(\gamma - a_1)^2}{2b_1^2}\right)\varphi_1(\theta) \tag{20}$$

$$\pi(\tau \mid \theta_{(\tau)}, X, Z) \propto \exp\left(-\frac{(\tau - a_2)^2}{2b_2^2}\right)\varphi_2(\theta) \tag{21}$$

onde

$$\varphi_1(\theta) = \exp\left(n_\bullet \gamma - \sum_{i=1}^{10} \sum_{l=1}^{n_i} \ln\left(1 + e^{\gamma + \tau w_i}\right) \right) \tag{22}$$

$$\varphi_2(\theta) = \exp\left(s\tau - \sum_{i=1}^{10} \sum_{l=1}^{n_i} \ln\left(1 + e^{\gamma + \tau w_i}\right) \right) \tag{23}$$

IG denota uma distribuição gama inversa, e, por exemplo, $\theta_{\left(\sigma_1^2\right)}$ é o vetor de todos os parâmetros, exceto σ_1^2. Notar que, como γ e τ não possuem distribuição condicional *a posteriori* conhecida, a geração de amostras para estes parâmetros necessita a utilização do algoritmo de Metropolis-Hastings (ver, por exemplo, Smith e Robert, 1993). A partir das distribuições condicionais *a posteriori* (14) a (21), o seguinte algoritmo é utilizado para gerar amostras das distribuições *a posteriori* dos parâmetros em θ:

(i) dado um valor inicial $\alpha_1^{(0)}, \alpha_2^{(0)}, \beta_1^{(0)}, \beta_2^{(0)}, \sigma_1^{2(0)}, \sigma_2^{2(0)}, \gamma^{(0)}$ e $\tau^{(0)}$, gerar $n_\bullet$ observações $Z_{il}^{(0)} (i = 1, \ldots, 10; l = 1, \ldots, n_i)$ da distribuição de Bernoulli com probabilidade de sucesso h_{il} dada em (9);

(ii) gerar $\alpha_1^{(1)}, \alpha_2^{(1)}, \beta_1^{(1)}, \beta_2^{(1)}, \sigma_1^{2(1)}, \sigma_2^{2(1)}, \gamma^{(1)}$ e $\tau^{(1)}$ a partir das distribuições condicionais *a posteriori* (14) a (21);

(iii) repetir (i) e (ii) até conseguir uma distribuição estacionária.

4. Resultados com os dados de Ribeirão Preto

As distribuições dos diâmetros das endurações observadas na população de Ribeirão Preto, em função do grupo etário, são apresentadas na Tabela 1 e na Figura 1. Os histogramas mostrados na Figura 1, referentes aos grupos etários até 24 anos, são semelhantes: elevada freqüência de reações 0-4 milímetros, que diminuem para indivíduos mais idosos, com quebra brusca até a coluna correspondente a 6 e 8 mm. A partir daí, as freqüências vão diminuindo de forma suave e gradativa até a extremidade direita onde estão as freqüências correspondentes às maiores reações. Observa-se uma nítida diferença entre os não-reatores até 3 mm e os demais, sendo a separação entre não-reatores e reatores situada na coluna correspondente aos diâmetros 4 e 5 mm.

A partir do grupo etário 25 a 29 anos, nota-se uma diminuição da proporção de não-reatores. Esboça-se a partir desse grupo etário uma tendência à distribuição bimodal, que se acentua com a idade, embora sem o aparecimento de uma nítida área de separação. Essa evolução dos histogramas sugere a ocorrência de destuberculização progressiva nessa população. Os gráficos obtidos nas pessoas mais velhas, representando situações epidemiológicas antigas, quando a prevalência de tuberculose era mais elevada.

Assim, para análise dos dados da Tabela 1, assumimos uma mistura de duas distribuições normais (1) com médias componentes dados por (3). Para a análise Bayesiana do modelo, substituímos os dados completos por uma amostra com 10% dos dados estratificados dentro de cada grupo etário e cada intervalo de reação em milímetros. Assim, por exemplo, consideramos 479 observações na faixa etária de 0 a 14 anos distribuídos proporcionalmente em cada intervalo de resultados do teste (reação em milímetros), isto é, 269 observações com reação no intervalo 0-1, 137 observações no intervalo 2-3, e assim sucessivamente. Dessa forma, temos uma amostra final de tamanho $n=3554$. A utilização de 10% dos dados na análise evita que números grandes sejam considerados nas funções exponenciais das equações (22) e (23), o que tornaria inviável o uso do algoritmo Bayesiano. Consideramos como covariável (w) o logaritmo do ponto médio da faixa etária em cada grupo.

Tabela 1 - Distribuição das reações tuberculínicas por diâmetro e por grupo etário

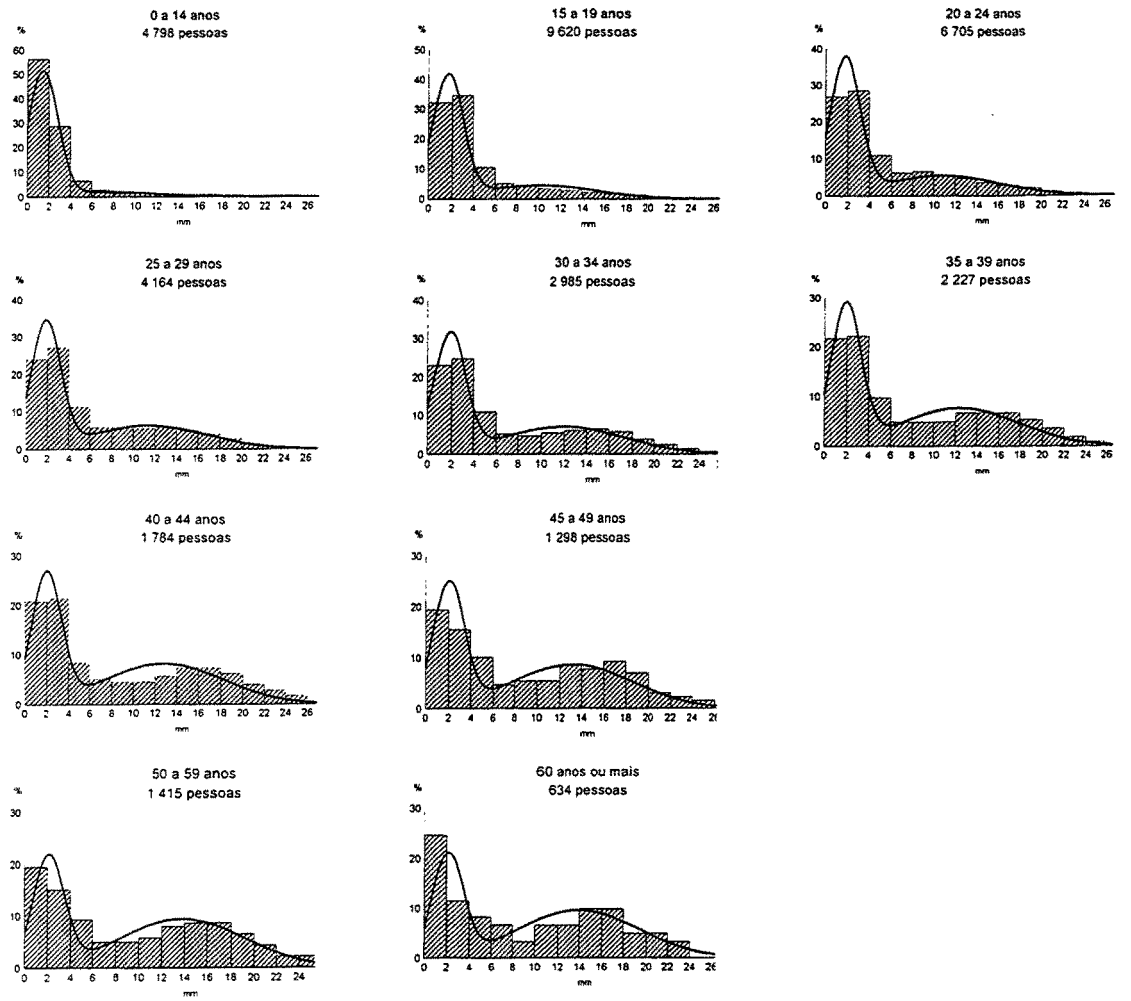
Reação Em	0 a 14 anos		15 a 19 anos		20 a 24 anos		25 a 29 anos		30 a 34 anos	
milímetros	*n*	%	*n*	%	*n*	%	*n*	%	*n*	%
0 - 1	2.688	56,03	3.097	32,19	1.804	26,91	1.001	24,04	678	22,71
2 - 3	1.368	28,52	3.337	34,69	1.913	28,54	1.135	27,26	734	24,59
4 - 5	312	6,50	1.016	10,56	728	10,86	469	11,26	319	10,69
6 - 7	133	2,77	492	5,11	421	6,28	239	5,74	154	5,16
8 - 9	62	1,29	420	4,37	443	6,61	240	5,76	142	4,76
10 - 11	65	1,35	344	3,58	359	5,35	230	5,52	163	5,46
12 - 13	48	1,00	268	2,79	316	4,71	237	5,69	177	5,93
14 - 15	36	0,75	206	2,14	237	3,53	214	5,14	192	6,43
16 - 17	38	0,79	159	1,65	190	2,83	170	4,08	168	5,63
18 - 19	32	0,67	128	1,33	139	2,07	116	2,79	115	3,85
20 - 21	10	0,21	80	0,83	83	1,24	60	1,44	74	2,48
22 - 23	3	0,06	45	0,47	39	0,58	26	0,62	36	1,21
24 - 25	1	0,02	16	0,17	20	0,30	14	0,34	21	0,70
26 - 27	0		9	0,09	4	0,06	4	0,10	5	0,17
28 - 29	2	0,04	3	0,03	7	0,10	5	0,12	4	0,13
30 ou mais	0		0		2	0,03	4	0,10	3	0,10
Total	4 798	100,0	9 620	100,0	6 705	100,0	4 161	100,0	2 985	100,0

Tabela 1 (cont.) - Distribuição das reações tuberculínicas por diâmetro e por grupo etário

Reação em milímetros	35 a 39 anos		40 a 44 anos		45 a 49 anos		50 a 59 anos		60 ou mais	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
0 - 1	492	21,61	375	21,02	249	19,18	274	19,37	152	23,97
2 - 3	504	22,13	380	21,30	202	15,56	214	15,13	71	11,20
4 - 5	222	9,75	154	8,63	130	10,02	127	8,98	49	7,73
6 - 7	108	4,74	90	5,04	64	4,93	68	4,81	37	5,84
8 - 9	114	5,01	79	4,43	68	5,24	71	5,02	25	3,94
10 - 11	113	4,96	84	4,71	68	5,24	84	5,94	37	5,84
12 - 13	146	6,41	99	5,55	109	8,40	111	7,84	44	6,94
14 - 15	147	6,46	134	7,51	105	8,09	119	8,41	57	8,99
16 - 17	153	6,72	135	7,57	118	9,09	123	8,69	65	10,25
18 - 19	124	5,45	107	6,00	87	6,70	94	6,64	33	5,21
20 - 21	78	3,43	69	3,87	44	3,39	62	4,38	27	4,26
22 - 23	39	1,71	47	2,63	27	2,08	34	2,40	20	3,15
24 - 25	21	0,98	19	1,07	13	1,00	19	1,34	6	0,95
26 - 27	6	0,26	8	0,45	8	0,62	7	0,49	5	0,79
28 - 29	6	0,26	2	0,11	2	0,15	3	0,21	3	0,47
30 ou mais	4	0,18	2	0,11	1	0,31	5	0,35	3	0,47
Total	2 277	100,0	1 784	100,0	1 298	100,0	1 415	100,0	634	100,0

Na análise Bayesiana destes dados, os valores dos hiperparâmetros das distribuições *a priori* para γ e τ foram escolhidos a partir de uma análise preliminar cuidadosa, que consistiu em considerar inicialmente o modelo (1) sem o efeito da covariável (w) nas probabilidades de incidência. A partir de estimativas preliminares para p_i, $i=1,\ldots,10$, foi possível escolher valores apropriados para os hiperparâmetros. Usando o algoritmo de amostradores de Gibbs conforme detalhado na seção anterior, geramos 100 000 amostras da distribuição *a posteriori* conjunta (13) a partir das distribuições condicionais *a posteriori* (14) a (21). Dessas 100 000 amostras, descartamos as 20 000 primeiras amostras para eliminarmos o efeito dos valores iniciais usados no algoritmo de simulação, e daí, consideramos a 20ª, 40ª , 60ª , ..., iterações, o que totaliza uma amostra final de tamanho 4 000 para a obtenção das inferências de interesse. O uso de uma observação a cada 20 garantiu a inexistência de correlações entre amostras seqüencialmente geradas.

Figura 1 - Distribuição dos diâmetros das provas tuberculínicas por grupo etário, ambos sexos, PPD rt 23, 2 U.T., Ribeirão Preto, 1973/74



A convergência do algoritmo foi verificada graficamente. Para a geração das amostras foi utilizado o software SAS. As Figuras 2 e 3 mostram graficamente o comportamento das cadeias finais, de tamanho 4 000, geradas pelo algoritmo, com seus respectivos histogramas ilustrando as formas das distribuições *a posteriori*.

As estimativas Bayesianas para os parâmetros do modelo são dadas na Tabela 2. Observa-se que a variância estimada para o segundo componente da mistura é bem maior que a estimada para o primeiro (σ_1^2 e σ_2^2 estimados em, respectivamente, 1,8386 e 29,2625). As estimativas dos parâmetros β_1 e β_2 são positivas (respectivamente 0,3167 e 3,1422), sugerindo que as médias

dos componentes na mistura sejam crescentes conforme as faixas etárias, sendo este crescimento mais acentuado para o segundo componente (dado que $\beta_2 > \beta_1$). Na Tabela 3, temos os sumários *a posteriori* das médias dos componentes da mistura em cada grupo etário. Estas médias foram estimadas a partir das amostras geradas para α_1, α_2, β_1 e β_2 (ver equação (3)) e dos pontos médios de cada faixa etária. A estimativa do parâmetro τ é negativo (−1,1456), sugerindo que a proporção p_1 do primeiro componente na mistura seja decrescente conforme avançam os grupos etários (Tabela 2). O intervalo de credibilidade 95% para τ exclui o valor zero, sugerindo que este decréscimo seja significativo. Na Tabela 4, os sumários *a posteriori* para as proporções do primeiro componente da mistura ilustram este decréscimo.

Tabela 2 - Estatísticas descritivas das distribuições *a posteriori* dos parâmetros do modelo

parâmetro	média	desvio padrão	intervalo de credibilidade 95%
α_1	-3,6023	0,0368	(-3,6736 ; -3,5326)
α_2	5,3274	0,2193	(4,8969 ; 5,7542)
β_1	0,3167	0,0573	(0,2056 ; 0,4299)
β_2	3,1422	0,3758	(2,3747 ; 3,8542)
σ_1^2	1,8386	0,0755	(1.6930 ; 1,9876)
σ_2^2	29,2625	1,3516	(26,6638 ; 31,9594)
γ	0,5311	0,0440	(0,4443 ; 0,6173)
τ	-1,1456	0,0891	(-1,3229 ; -0,9724)

Tabela 3 - Estatísticas descritivas das distribuições *a posteriori* das médias dos componentes na mistura

Grupo etário	média do componente 1 na mistura (μ_{1i})			média do componente 2 na mistura (μ_{2i})		
	média	Desvio padrão	intervalo de credibilidade 95%	média	Desvio padrão	intervalo de credibilidade 95%
0 a 14	1,4412	0,0639	(1,3162 ; 1,5631)	7,1974	0,5419	(6,1562 ; 8,2705)
15 a 19	1,7223	0,0352	(1,6540 ; 1,7901)	9,9854	0,2647	(9,4576 ; 10,5063)
20 a 24	1,8039	0,0372	(1,7317 ; 1,8746)	10,7956	0,2164	(10,3720 ; 11,2179)
25 a 29	1,8688	0,0425	(1,7839 ; 1,9501)	11,4391	0,2043	(11,0404 ; 11,8364)
30 a 34	1,9226	0,0487	(1,8238 ; 2,0199)	11,9729	0,2156	(11,5527 ; 12,3905)
35 a 39	1,9686	0,0548	(1,8566 ; 2,0780)	12,4291	0,2388	(11,9609 ; 12,8968)
40 a 44	2,0087	0,0605	(1,8863 ; 2,1287)	12,8274	0,2666	(12,2968 ; 13,3486)
45 a 49	2,0443	0,0659	(1,9112 ; 2,1745)	13,1808	0,2956	(12,5959 ; 13,7693)
50 a 59	2,1054	0,0755	(1,9547 ; 2,2546)	13,7870	0,3516	(13,0809 ; 14,4722)
60 ou mais	2,1217	0,0781	(1,9659 ; 2,2761)	13,9481	0,3674	(13,2080 ; 14,6630)

Tabela 4 - Estatísticas descritivas das distribuições *a posteriori* das proporções do primeiro componente da mistura

grupo etário	proporção do componente 1 na mistura (p_{1i})		
	média	desvio padrão	intervalo de credibilidade 95%
0 a 14	0,8597	0,0141	(0,8302 ; 0,8863)
15 a 19	0,6901	0,0112	(0,6675 ; 0,7114)
20 a 24	0,6238	0,0102	(0,6035 ; 0,6438)
25 a 29	0,5674	0,0110	(0,5457 ; 0,5890)
30 a 34	0,5192	0,0127	(0,4937 ; 0,5439)
35 a 39	0,4776	0,0146	(0,4483 ; 0,5062)
40 a 44	0,4416	0,0165	(0,4093 ; 0,4737)
45 a 49	0,4102	0,0181	(0,3744 ; 0,4453)
50 a 59	0,3581	0,0205	(0,3183 ; 0,3977)
60 ou mais	0,3448	0,0211	(0,3041 ; 0,3855)

Na Figura 1, temos as curvas ajustadas pelo modelo sobrepostas aos histogramas, onde observamos um bom ajuste do modelo aos dados de tuberculose da Tabela 1. Sendo o envelhecimento da população uma das principais causas para a gravidade da situação atual da tuberculose no mundo Ruffino-Netto (2002), o modelo proposto aplicado a dados atualizados poderá auxiliar a compreender o impacto de situações epidemiológicas pregressas sobre a prevalência atual da doença.

Figura 2 – Comportamento das cadeias geradas e histogramas *a posteriori*, para os parâmetros α_1 ((a) e (b)), α_2 ((c) e (d)), β_1 ((e) e (f)) e β_2 ((g) e (h))

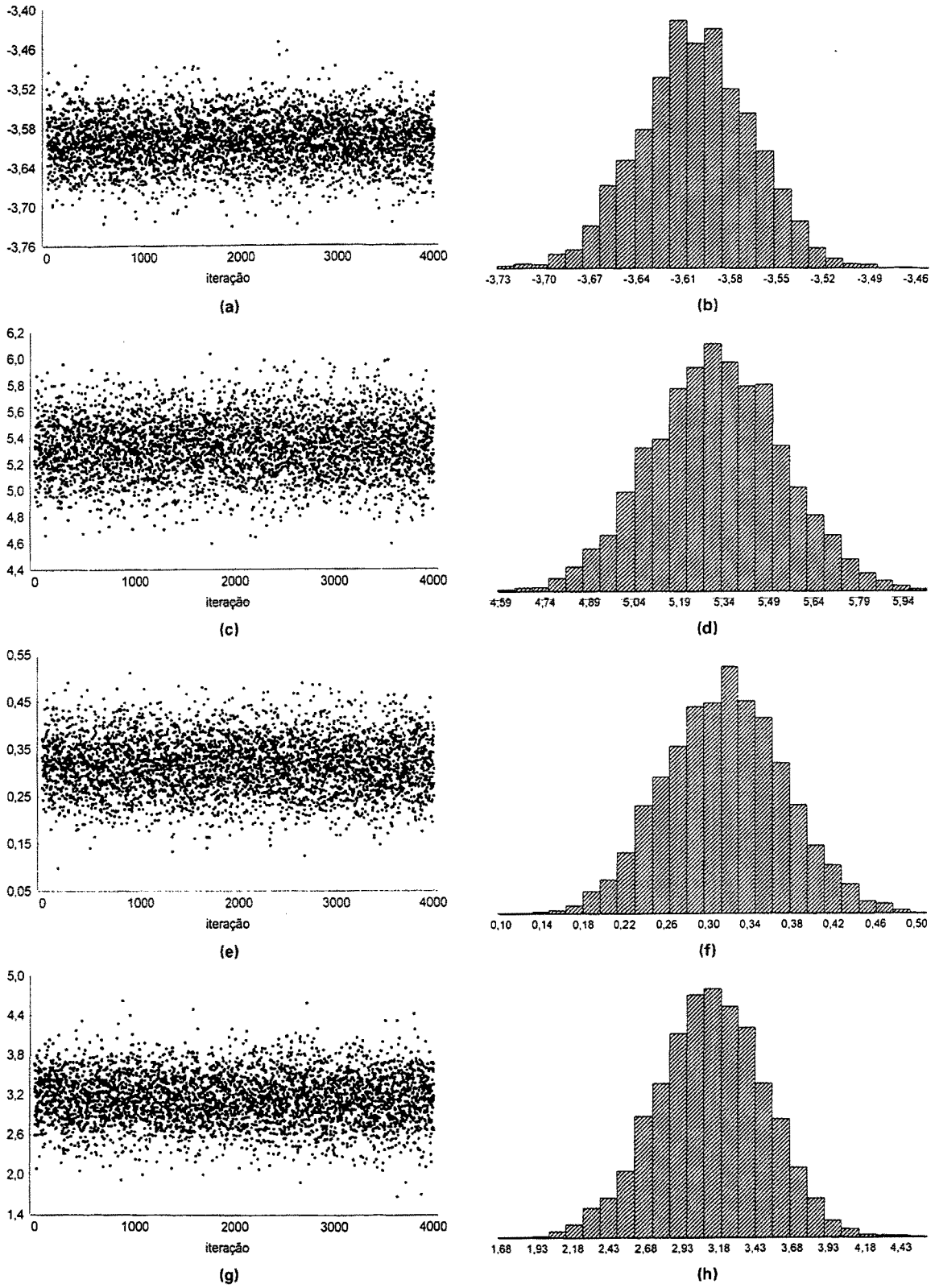
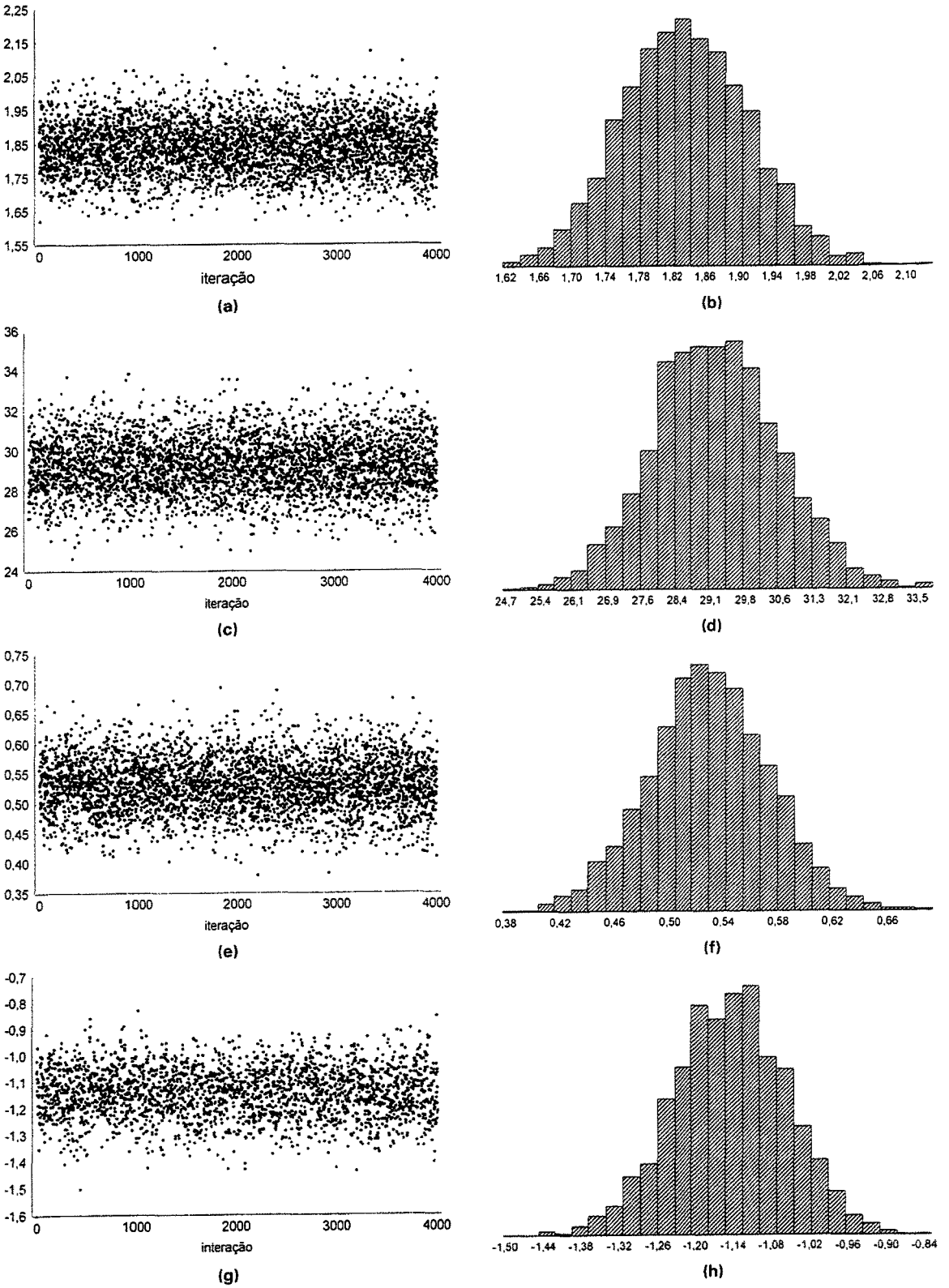


Figura 3 – Comportamento das cadeias geradas e histogramas *a posteriori*, para os parâmetros σ_1^2 ((a) e (b)), σ_2^2 ((c) e (d)), γ ((e) e (f)) e τ ((g) e (h))



5. Conclusões

O presente estudo apresentou uma eficiente alternativa de estimação de parâmetros de um modelo de mistura de distribuições normais. Sua aplicação aos dados reais descreveu satisfatoriamente como os resultados da prova tuberculínica se distribuem de acordo com as faixas etárias, captando as diferenças entre as duas componentes da mistura de distribuições. No modelo apresentado, o uso de variâncias constantes σ_1^2 e σ_2^2 para todas as faixas etárias pode ser visto como uma limitação da proposta. Entretanto, embora possível, a inserção de um componente linear para descrever possíveis mudanças das variâncias conforme uma covariável teria como prejuízo a perda da simplicidade do modelo apresentado, o que extrapola o objetivo deste estudo.

Embora o presente estudo tenha como motivação a análise de resultados da prova tuberculínica, o modelo apresentado pode ser adaptado a outras situações de interesse.

Referências bibliográficas

WHO - World Health Organization. WHO Report 2003 – Global Tuberculosis Control – Surveillance, Planning, Financing. WHO/CDS/TB, 2003.

CARNEIRO, JF. Aspectos atuais da alergia na tuberculose. *Revista do Serviço Nacional de Tuberculose* 1962; 6: 305-360.

RUFFINO-NETO A, SANCHES O, ARANTES GR. Aplicação do método de Bhattacharya na análise de resultados do teste tuberculínico. *Rev. Saúde Pública* 1977; 11: 322-329.

COHEN AC. Estimation in mixtures of two normal distributions. *Technometrics* 1967; 9: 15-28.

DAY NE. Estimating the components of a mixture of normal distributions. *Biometrika* 1969; 56: 463-474.

DICK NP, BOWDEN DC. Maximum likelihood estimation for mixtures of two normal distributions. *Biometrics* 1973; 29: 781-790.

NEAL RM. *Bayesian mixture modeling by Monte Carlo simulation.* Technical report, Department of Computer Science. Toronto: University of Toronto, 1991.

DIEBOLT J, ROBERT CP. Estimation of finite mixture distributions through Bayesian sampling. *J. R. Statist. Soc. B* 1994; 56:363-375.

ESCOBAR MD, West M. Bayesian density estimation and inference using mixtures. *J. Am. Statist. Ass.* 1995; 90:577-588.

RICHARDSON S, GREEN PJ. On Bayesian analysis of mixtures with an unknown number of components. *J. R. Statist. Soc. B* 1997; 59:731-792.

ROBERTS SJ, HUSMEIER D, REZEK I, PENNY W. Bayesian approaches to Gaussian mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* 1998; 20:887-906.

TITTERINGTON DM, SMITH AFM, MAKOV UE. *Statistical analysis of finite mixture distributions.* New York: John Wiley, 1985.

TANNER M, WONG W. The calculation of posterior distributions by data augmentation. *J. Am. Statist. Ass.* 1987; 82: 528-550.

BOX GEP, TIAO GC. *Bayesian inference in statistical analysis.* Reading: Addison-Wesley, 1973.

GELFAND AE, SMITH AFM. Sampling based approaches to calculating marginal densities. *J. Am. Statist. Ass.* 1990; 85: 398-409.

SMITH AFM, ROBERT GO. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc. B* 1993; 55: 3-24.

RUFFINO-NETTO A. Tuberculosis: the neglected calamity. *Rev. Soc. Bras. Med. Trop.* 2002; 35: 51-58.

Abstract

The distribution of the tuberculin reactions diameter is usually given by a mixture of two or more probability distributions. In this paper, considering the results of a tuberculin test data set from Ribeirão Preto, we considered the use of a Bayesian model that assumes a mixture of normal distributions for the tuberculin reactions diameters. When compared with other approaches, the model proposed has the advantage of easily implementation and allows the inclusion of one or more covariates. We used simulation methods based in MCMC algorithms (Markov chain Monte Carlo) to get the Bayesian summaries of interest. The model allowed describing the tuberculin test results distribution according to age groups.

Key words: tuberculosis, tuberculin test, bayesian methods.

Modelos de resposta aleatória na avaliação de itens sensíveis: a utilização de drogas ilícitas por alunos de graduação na Universidade de Brasília

Fernanda Gomes Philomeno*
Priscila dos Santos Fonseca**
Geraldo da Silva e Souza***

Resumo

Este artigo utiliza amostragem estratificada no contexto de modelos de resposta aleatória para avaliar a intensidade da utilização de drogas ilícitas pelo corpo discente de graduação da Universidade de Brasília. A estratificação levada a efeito é definida pelas áreas de Ciências Humanas, Ciências Exatas e Ciências Biológicas. Estimativas da proporção de usuários de drogas são obtidas por estrato e para a população como um todo. A classificação por sexo é investigada como um domínio de estudo. Os resultados globais obtidos são socialmente preocupantes. Estima-se em 25% (±4,1) a proporção de usuários. Para usuários do sexo feminino a intensidade é 20,2% (±6,0) e para usuários do sexo masculino obtém-se 29,0% (±5,8). As estimativas pontuais para os estratos Humanas, Exatas e Biológicas são, respectivamente, 28,1% (±5,6), 23,1% (±4,4) e 18,8% (±8,7). A análise da amostra via regressão logística indica diferenças significantes apenas entre os sexos.

Palavras-Chave: Amostragem aleatória estratificada, modelos de resposta aleatória, uso de drogas ilícitas.

* Endereços para correspondências: Departamento de Estatística, Universidade de Brasília, e-mail: fephi@pop.com.br.
** e-mail: priscila.fonseca@uol.com.br.
*** e-mail: geraldosouza@unb.br.

1. Introdução

Neste artigo aplica-se a técnica de amostragem estratificada, com a utilização de modelos de resposta aleatória, para avaliar a intensidade do uso de drogas ilícitas pelo corpo discente da Universidade de Brasília. A população alvo investigada é definida pelos alunos de graduação matriculados no segundo semestre de 2003 e se constitui de 19.366 estudantes.

A estratificação adotada é definida pelas áreas de Ciências Humanas (Administração, Arquitetura e Urbanismo, Artes Cênicas, Artes Plásticas, Biblioteconomia, Ciência Política, Ciências Contábeis, Ciências Econômicas, Ciências Sociais, Comunicação Social, Desenho Industrial, Direito, Educação Artística, Filosofia, Geografia, História, Letras, Música, Pedagogia, Relações Internacionais, Serviço Social e Arquivologia), Ciências Exatas (Ciência da Computação, Engenharia Civil e Ambiental, Engenharia Elétrica, Engenharia Mecânica, Estatística, Física, Geologia e Matemática) e Ciências Biológicas (Agronomia, Biologia, Educação Física, Enfermagem e Obstetrícia, Engenharia Florestal, Farmácia, Medicina, Veterinária, Nutrição, Odontologia e Psicologia). Esses estratos são compostos de, respectivamente, 10 899, 4 469 e 3 998 estudantes.

O conceito de droga ilícita utilizado neste artigo é definido por substâncias para as quais se considera a produção, a comercialização e o uso como crimes. Os produtos considerados aqui foram maconha, haxixe, cocaína, ópio, heroína, craque, mesclado e merla.

Os modelos de resposta aleatória foram introduzidos na literatura estatística por Warner (1965) e representam uma abordagem engenhosa para lidar com questões delicadas ou sensíveis em pesquisas amostrais. Por itens sensíveis entende-se questões cuja resposta pode causar constrangimento ao entrevistado. Exemplos são dados pela questão associada a este artigo "Você faz uso de droga ilícita pelo menos uma vez por semana?" e perguntas como "Você sonega o imposto de renda?" ou "Você é homossexual?". Em tais casos a pergunta direta, tipicamente, leva a não resposta ou a resposta falsa. A idéia então é obter informação confiável de forma indireta, com viés menor, a partir de um mecanismo de aleatorização que evita que a classe a que pertence o respondente seja conhecida pelo entrevistador.

O modelo de resposta aleatória binomial original de Warner (1965) e a extensão de Horvitz, Shah e Simmons (1967), que usaremos aqui, conhecida como modelo de resposta aleatoria com questão não relacionada, estão descritos em Cochran (1977), Scheaffer, Mendenhall e Ott (1996) e Lohr (1999). Tais modelos admitem generalizações para questões multinomiais e de natureza quantitativa. Para maiores detalhes veja Philomeno e Fonseca (2003). Uma aplicação

recente do modelo de resposta aleatória, num contexto semelhante ao aqui abordado, é apresentada em Mauer et al. (2004).

Nossa exposição no artigo procede como segue. Na seção 2 introduzimos os modelos de resposta aleatória tais como apresentados em Cochran (1977) e em Philomeno e Fonseca (2003). Na seção 3 apresentam-se os detalhes técnicos do programa amostral – questões de tamanho de amostra e de técnicas de estimação, e as estatísticas resultantes do programa amostral levado a efeito na Universidade de Brasília. Na seção 4 apresentam-se e comentam-se as estimativas obtidas. Finalmente na seção 5 apresenta-se um resumo final do trabalho com conclusões.

2. Modelos de resposta aleatória

Considere uma população humana composta de N indivíduos. Estamos interessados em determinar a proporção π de indivíduos nesta população que possuem a característica sensível aqui representada por A. Warner (1965) introduziu a técnica de resposta aleatória para estimar π. O método tem por objetivo eliminar a incidência de não resposta e de respostas falsas na pesquisa. Warner (1965) mostrou que, com o uso de um mecanismo de aleatorização, é possível estimar π de forma não viesada e sem que os respondentes revelem diretamente sua classificação quanto a presença de A (e deste modo preservando a confidência da resposta).

Suponha que um elemento típico da população seja convidado a responder Sim ou Não a uma das afirmações seguintes:

1. Eu tenho a característica A.
2. Eu não tenho a característica A.

Suponha adicionalmente que com probabilidade P conhecida, e distinta de 1/2, o respondente responde a afirmação (1) e com probabilidade 1 - P responde a afirmação (2). Para uma amostra de tamanho n o investigador calcula a proporção $\hat{\lambda}$ de respostas Sim na amostra, sem saber se cada Sim é, individualmente, uma manifestação relativa à afirmação (1) ou (2).

Se λ *representa a probabilidade de resposta afirmativa, então a relação entre* π, *a* probabilidade do indivíduo possuir a caraterística A, e λ, é dada por

$$\lambda = P\pi + (1-P)(1-\pi) = (2P-1)\pi + (1-P) \qquad (1)$$

Como λ pode ser estimado eficientemente por $\hat{\lambda}$, a Equação 1 sugere como estimador de π a expressão

$$\hat{\pi} = \frac{\hat{\lambda} - (1-P)}{2P-1}. \tag{2}$$

Num esquema de amostragem aleatória simples o estimador dado pela Equação 2 é não viesado e tem variância

$$Var(\hat{\pi}) = \left(\frac{N-n}{N-1}\right)\frac{\lambda(1-\lambda)}{(2P-1)^2}. \tag{3}$$

Sob a hipótese de amostragem aleatória de uma população infinita, i.e., se as observações amostrais forem vistas como amostras independentes e identicamente distribuídas (iid) de uma população Bernoulli com probabilidade $\lambda = \pi P(+1-\pi)(1-P)$ de sucesso então a função de verossimilhança do ensaio amostral é dada por

$$L(x,\pi) = \left[\pi P + (1-\pi)(1-P)\right]^x \left[(1-\pi)P + \pi(1-P)\right]^{(n-x)} \tag{4}$$

onde x é o número de sucessos (Sim's) observados na amostra. O estimador $\hat{\pi}$ maximiza a função de verossimilhança descrita na Equação 4.

Um método alternativo ao modelo de Warner (1965) foi sugerido por Horvitz, Shah e Simmons (1967). Estes autores perceberam que dando oportunidade ao entrevistado de responder uma de duas questões, na qual uma é completamente inócua e não relacionada com o atributo estigmatizado, o respondente poderia ser mais sincero.

No modelo de Horvitz, Shah e Simmons (1967) o indivíduo responde a uma das questões seguintes, com probabilidades conhecidas P e $1-P$, respectivamente, sem revelar ao entrevistador que questão está respondendo.

1. Você possui a característica A? (pergunta sensível)
2. Você pertence à categoria Y? (pergunta não relacionada)

Note que a pergunta não relacionada não causa nenhum constrangimento ao entrevistado. Supõe-se conhecida a proporção π_y de indivíduos na população que pertencem à categoria Y.[1]

[1] O procedimento de estimação também pode ser levado a efeito se π_y não for conhecido. Veja Philomeno e Fonseca (2003).

A relação entre λ e π na população é dada por

$$\lambda = P\pi + (1-P)\pi_y . \tag{5}$$

Obtém-se da Equação 5 o estimador

$$\hat{\pi} = \frac{\hat{\lambda} - \pi_y(1-P)}{P} \tag{6}$$

com variância, no esquema de amostragem aleatória simples,

$$Var(\hat{\pi}) = \left(\frac{N-n}{N-1}\right)\frac{\lambda(1-\lambda)}{P^2} . \tag{7}$$

Sob a hipótese de amostragem aleatória iid de uma população Bernoulli com probabilidade λ de sucesso, tem-se

$$L(x,\pi) = \left[P(\pi - \pi_y) + \pi_y\right]^x \left[P(\pi_y - \pi_u) + (1-\pi_y)\right]^{(n-x)} \tag{8}$$

sendo x o número de sucessos (Sim's). Vê-se que $\hat{\pi}$ maximiza a função de verossimilhança dada na Equação 8. Note que não há restrições quanto ao valor de $P \in (0,1)$.

Como $(2P-1)^2 < P^2$ para $1/3 < P < 1$, a variância do estimador dado pela Equação 6 é estritamente menor do que a do estimador de Warner que é dado pela Equação 2. As escolhas usuais de P são feitas no intervalo [0,5; 1]. Obviamente a variância diminui com P e, portanto, quanto maior P, mais informativa a amostra. Com P muito perto de um, contudo, corre-se o risco dos respondentes acharem que o entrevistador sabe qual questão está sendo respondida. É claro também que o valor de P influi no processo de escolha do tamanho da amostra. Como veremos na seção 3, para $\lambda = 0{,}275$, se limitarmos o erro máximo na estimativa de π em 3%, para mais ou para menos, a escolha $P = 0{,}5$ exige uma amostra de tamanho 1742. A escolha $P = 0{,}75$ requer uma amostra de tamanho 816. O valor de P sugerido na literatura é $P = 0{,}75$ (Scheaffer, Mendenhall e Ott, 1996).

O modelo proposto por Horvitz, Shah e Simmons (1967) pode ser facilmente implementado, em aplicações, com a utilização do método sugerido em Greenberg et al. (1969). Este é como segue. Toma-se como dispositivo de aleatorização uma caixa contendo bolas vermelhas, brancas e azuis nas proporções P_1, P_2 e P_3 respectivamente. O respondente escolhe

uma bola ao acaso e segue as instruções de acordo com sua cor:

1. Bola vermelha: Responda a pergunta: Você possui a característica (sensível) A?
2. Bola de cor azul ou branca: Responda a pergunta: A bola escolhida tem cor branca?

Neste caso tem-se $P = P_1, \pi_y = P_2/(P_2 + P_3)$ e as Equações 5 e 6 se tornam $\lambda = P_1\pi + P_2$ e $\hat{\pi} = (\hat{\lambda} - P_2)/P$, respectivamente.

3. Plano amostral

Vê-se da Equação 6 que para estimar π tudo que se necessita é uma estimativa de λ. Como

$$Var(\hat{\pi}) = Var(\hat{\lambda})/P^2 \tag{9}$$

o erro B_π na estimativa de π, com nível de confiança $1-\alpha$, $\alpha \in (0,1)$, será obtido se considerarmos o erro $B = P \times B_\pi$, na estimativa de λ, com o mesmo nível probabilístico $1-\alpha$. Deste modo o parâmetro λ torna-se, na realidade, a característica de interesse primário do programa amostral.

O planejamento amostral adotado no presente trabalho para a obtenção de uma estimativa de λ foi o de amostragem estratificada com alocação proporcional da amostra. Neste contexto seja N_i, $N_1 = 10.899$, $N_2 = 4.469$, $N_3 = 3.998$, a população do estrato $i = 1$ *(Humanas), 2 (Exatas), 3 (Biológicas)* e $W_i = N_i/N$, $N = 19.366$, $W_1 = 0{,}5628$, $W_2 = 0{,}2308$, $W_3 = 0{,}2064$.

A variância do estimador $\hat{\lambda}$ (da proporção λ),

$$\hat{\lambda} = \sum_{i=1}^{3} W_i \hat{\lambda}_i \tag{10}$$

é dada por

$$Var(\hat{\lambda}) = \frac{(1-f)}{n} \sum_{i=1}^{3} W_i S_i^2 \tag{11}$$

com

$$S_i^2 = \frac{N_i}{N_i - 1}\lambda_i(1-\lambda_i) \tag{12}$$

sendo λ_i a freqüência relativa de Sim's no estrato i, $f = n/N$ e $\hat{\lambda}_i$ a proporção amostral correspondente.

A classificação sexo foi tratada como um domínio de estudo (Cochran, 1977, Seção 2.12) e não como definidora de estratos.

Supondo o mesmo valor de λ em cada estrato, o tamanho da amostra necessário para se obter um erro na estimativa de λ de no máximo B, para mais ou para menos, com probabilidade $1-\alpha$, é

$$\frac{\lambda(1-\lambda)}{V^* + \lambda(1-\lambda)/N} \tag{13}$$

onde $V^* = B^2 / z_\alpha^2$ e $\phi(z_\alpha) = 1-\alpha/2$. Aqui $\phi(.)$ representa a função de distribuição da normal padrão. O erro máximo associado à estimativa de π, para o mesmo nível probabilístico $1-\alpha$, é B/P.

A fórmula do tamanho da amostra, dada pela Equação 13, depende de uma estimativa preliminar de λ. A escolha $\lambda = 0,5$ é muito conservativa e conduz a valores muito elevados para o tamanho da amostra. Com o intuito de obter uma estimativa preliminar de λ menos conservativa e, simultaneamente, testar os instrumentos da pesquisa, notadamente o mecanismo de aleatorização, tomou-se uma amostra de estudantes do Departamento de Estatística. Neste programa piloto a amostra foi composta de 40 estudantes (de uma população total de 143 estudantes de graduação matriculados no primeiro semestre de 2003). O modelo de resposta aleatória escolhido foi o de Horvitz, Shah e Simmons (1967). Como instrumento de aleatorização tomou-se um baralho contendo 30 cartas de copas, 5 cartas de espadas e 5 cartas de paus. As cartas de copas, paus e espadas têm o mesmo papel das bolas vermelhas, brancas e azuis da técnica sugerida por Greenberg et al. (1969). A pergunta sensível, a mesma da pesquisa principal, foi "Você faz uso de drogas ilícitas pelo menos uma vez por semana?". Esta pergunta deveria ser respondida caso a carta escolhida fosse de copas. A questão não relacionada foi "A carta sorteada é de paus?". Tem-se portanto $P = 3/4$ e $\pi_y = 1/2$.

Para o ensaio piloto obtivemos $\hat{\lambda} = 0{,}275$. Este valor foi considerado como aproximação de λ na Equação 13.

A Tabela 1 mostra as alternativas de tamanho de amostra consideradas preliminarmente, todas com nível probabilístico de 95%.

Nossa escolha recaiu sobre a alternativa $n = 816$, correspondente a um erro esperado de 3% para mais ou para menos na estimativa global de π. Erros menores do que este são impraticáveis face a limitação de recursos e maiores colocam em risco a representatividade do trabalho.

Faz-se mister observar aqui que, de fato, o nível máximo de erro estimado, a partir da amostra, foi de 4,1%.

Tabela 1 - Alternativas de erros e tamanhos de amostra

Erro em $\hat{\lambda}(B)(\%)$	Erro em $\hat{\pi}(B/P)(\%)$	n
0,75	1	5.489
1,50	2	1.742
2,25	3	816
3,00	4	467
3,75	5	302

4. Resultados amostrais e análise estatística

Os resultados amostrais encontrados estão resumidos na Tabela 2. Ali apresentam-se a distribuição das repostas Sim's e Não's pelas classificações geradas pela interação de estrato e sexo. Diferenças marcantes de respostas Sim e Não para as várias classificações induzidas pelas classificações de estrato e sexo são indicativas de diferenças na proporção de uso de drogas por essas populações, uma vez que as primeiras caracterizam as últimas.

As estimativas obtidas com o uso do modelo de resposta aleatória e amostragem estratificada são apresentadas na Tabela 3.

Tabela 2 - Resultados amostrais

Estrato	Sexo				
	Masculino		Feminino		
	Sim	Não	Sim	Não	Total
Humanas	88	148	66	157	459
Exatas	46	98	10	34	188
Biológicas	19	44	26	80	169

Tabela 3 - Proporções (%) estimadas de usários de drogas ilícitas (Valores entre parêntesis representam desvios padrão)

Estrato	Sexo		Total	Intervalo a 95%
	Masculino	Feminino		
Humanas	33,1 (4,1)	22,8 (4,0)	28,1 (2,9)	[22,4:33,8]
Exatas	25,9 (5,1)	13,6 (8,3)	23,1 (4,4)	[14,5:31,7]
Biológicas	23,5 (7,6)	16,0 (5,5)	18,8 (4,5)	[10,0:27,6]
Total	29,0 (3,0)	20,2 (3,1)	25,0 (2,1)	[20,9:29,1]

Observa-se na Tabela 3 que os intervalos com coeficiente de 95% para as proporções de uso nos estratos se interceptam. O mesmo ocorre com os intervalos associados às proporções de uso por sexo, não mostrados na Tabela 3. Isto é indicativo de que não existem diferenças significantes entre sexos e entre áreas do conhecimento. Procuramos a confirmação desses resultados com o uso de regressão logística.[2]

A regressão logística, associada à modelagem de λ como função dos efeitos estrato, sexo e da interação estrato versus sexo, produziu os resultados constantes da Tabela 4.

[2] As Tabelas 4-6 foram obtidas com o uso do SAS (*Statistical Analysis System*).

Tabela 4 - Regressão Logística com interação estrato x sexo

Efeito	gl	Qui-quadrado - Wald	Pr > Qui-quadrado
Estrato	2	1,7485	0,4172
Sexo	1	3,0328	0,0816
Estrato x Sexo	2	0,11193	0,9421
-2 Log L	6	1.005,469	

Na Tabela 4 vê-se que a interação estrato versus sexo não é significante. Deste modo considerou-se um modelo reduzido só com efeitos principais. As Tabelas 5 e 6 mostram os resultados do ajuste deste modelo.

Tabela 5 - Regressão Logística com Efeitos Principais

Efeito	gl	Qui-quadrado - Wald	Pr > Qui-quadrado
Estrato	2	3,0971	0,2126
Sexo	1	4,9456	0,0262
-2 Log L	4	1.005,590	

Vê-se da Tabela 5 que o teste da razão de verossimilhança da interação está de acordo com o teste de Wald e que somente sexo tem efeito significante. O sexo masculino apresenta intensidade maior de consumo.

As Estatísticas de bondade do ajuste mostradas na Tabela 6 são indicativas de que as hipóteses do modelo logístico são aceitáveis.

Tabela 6 - Estatísticas de bondade do ajuste logístico - Modelo Reduzido

Estatística	gl	Valor	Valor/gl	Pr > Qui-quadrado
Deviance	2	0,1207	0,0604	0,9414
Pearson χ^2	2	0,1197	0,0599	0,9419

5. Resumo e Conclusões

Levou-se a efeito um programa amostral com o objetivo de se estimar a proporção de usuários de drogas ilícitas na população de estudantes de graduação da Universidade de Brasília matriculados no segundo semestre de 2003. Por droga ilícita entende-se agentes cuja produção, comercialização e uso são considerados como crime (maconha, haxixe, cocaína, ópio, heroína, craque, merla, mesclado e ecstasy). A pergunta de interesse da pesquisa foi "Você faz uso de droga ilícita pelo menos uma vez por semana?".

A população alvo se constitui de 19.366 estudantes dos quais foram entrevistados 816. O plano amostral utilizado foi de amostragem estratificada com alocação proporcional, no contexto de modelos de resposta aleatória. As subpopulações de interesse foram determinadas pela interação de estratos (3 níveis - Ciências Humanas, Exatas e Biológicas) com sexo (2 níveis - masculino e feminino) e o modelo de resposta aleatória utilizado foi o de Horvitz, Shah e Simmons (1967), que envolve uma questão não relacionada à questão de interesse da pesquisa (questão sensível). Cada respondente foi submetido a um mecanismo de aleatorização que com probabilidade 3/4 produz uma resposta associada à questão sensível . As probabilidades de respostas afirmativa e negativa para a pergunta não relacionada foram iguais a 0,5.

Os resultados globais obtidos para a pesquisa, indicam índices elevados de consumo de drogas, tanto para a população, como para as subclassificações de interesse - áreas de conhecimento e sexo. O percentual de uso de drogas estimado para a população foi de 25% ($\pm$4,1).

Para a população masculina o valor obtido foi 29% ($\pm$5,8) e para a população feminina 20,2% ($\pm$6,0). Os valores obtidos para as áreas de Ciências Humanas, Exatas e Biológicas foram 28,1% ($\pm$5,6), 23,1% ($\pm$4,4) e 18,8% ($\pm$8,7), respectivamente. Esses resultados sugerem um cenário de uso intenso de drogas pelos alunos componentes da população objetivo.

Uma análise com o uso de regressão logística tendo como base os dados da pesquisa indica a presença de efeitos significantes entre os níveis de sexo. Não se observam diferenças significantes entre os perfis dos estratos nem de interação entre estrato e sexo.

A utilização dos modelos de resposta aleatória em planejamentos amostrais demandam tamanhos de amostra relativamente grandes para que se obtenha um nível aceitável de erro.

A abordagem se justifica, no entanto, pela eliminação dos vieses causados por não resposta e respostas falsas. Tal é o caso da presente aplicação em que o item sob investigação é evidentemente sensível. Vale ressaltar contudo que o uso de drogas está se tornando um lugar

comum na UnB. Evidência disto é que vários estudantes quando entrevistados se dispunham a fornecer respostas diretas, dispensando a utilização do mecanismo de aleatorização, sem o menor constrangimento.

Referências bibliográficas

COCHRAN, W. G. (1977). *Sampling Techniques*. 3rd. ed., Wiley (New York).

GREENBERG, B. G.; ABUL-ELA, A. A.; SIMMONS, W. R.; HORVITZ, D. G. (1969). "The Unrelated Randomized Response Model: Theoretical Framework". *Journal of the American Statistical Association*, 64, 520-539.

HORVITZ, D. G.; SHAH, B. C.; SIMMONS, W. R. (1967). "The Unrelated Randomized response Model". *Proc. Soc. Stat. American Statistical Association*, 65-72.

LOHR, S. L. (1999). *Sampling Design*. Duxbury (New York).

MAUER, M.; DONNEAU, A.; PASQUASY, N.; REGGERS, J.; GOSSET, C.; ALBERT, A.; SARTOR, F. (2004). "Random Response Model for Estimating Illicit Drug Prevalence among Youth. A Feasibility Study". Report IPH/EPI 2004-012, Brussels' Scientific Institute of Public Health, Epidemiology Unit, Belgium.

PHILOMENO, F. G.; FONSECA, P. S. (2003). "Modelos de Resposta Aleatória para Itens Sensíveis". Relatório de estágio supervisiosnado, Departamento de Estatística, Universidade de Brasília.

SCHEAFFER, R. L.; MENDENHALL, W.; OTT, R. L. (1996). *Elementary Survey Sampling*. 5th ed., Duxbury (New York).

WARNER, S. L. (1965). "Randomized Response: A Survey Technique for eliminating Evasive Answer Bias". *Journal of the American Statistical Association*, 60, 63-69.

Abstract

This article uses stratified random sampling in the context of randomized response models to assess the intensity of illicit drug use by undergraduate students enrolled in the University of Brasília. Stratification is defined by the areas of Human Sciences, Exact Sciences and Biological Sciences. Estimates of proportions of drug users are computed for each stratum and for the whole population. The classification according to sex is investigated as a domain of study. The overall results are of social concern. The proportion of drug users is estimated in 25% (±4,1). For women the intensity of use is 20,2% (±6,0) and for men is 29,0% (±5,8). Point estimates for the strata Human Sciences, Exact Sciences, and Biological Sciences are, respectively, 28,1% (±5,6), 23,1% (±4,4), and 18,8% (±8,7). The analysis of the sampling data via logistic regression indicates significant di_erences only between sex levels.

Key Words: Stratified Random Sampling , Randomized Response Models, Use of Illicit Drugs.

Modelos de fronteira de produção: a eficiência dos institutos da UFRJ

Viviane C. C. Quintaes*
Fernando A. S. Moura**
Hélio S. Migon ***

Resumo

Neste estudo uma revisão de modelos de fronteira de produção estocástica é apresentada. Dentre os problemas abordados, destacam-se: a escolha da distribuição para representar as eficiências, a modelagem de dados de painel e as alternativas metodológicas disponíveis para o tratamento de situações com múltiplos produtos. Os modelos introduzidos são membros da classe dos modelos hierárquicos Bayesianos e as inferências serão realizadas através de métodos de simulação estocástica - Markov Chain Monte Carlo. Assim os efeitos relativos dos insumos sobre a produção -elasticidades- podem ser modeladas como firma específica ou não.

Os modelos revisados são aplicados a dados reais relativos a Universidade Federal do Rio de Janeiro - UFRJ e avaliados em três períodos de tempo. Os resultados obtidos serão comentados e ilustrados, destacando-se as potencialidades dessas técnicas como ferramentas úteis na tomada de decisão.

* Endereço par correspondência: COPPEUFRJ, e-mail: vquintaes@hotmail.com
** IMUFRJ, e-mail: fmoura@im.ufrj .br
*** IM/COPPEUFRJ, e-mail: migon@im.ufrj.br

1. Introdução

A partir da especificação de uma função de produção que retrate a tecnologia disponível, pode-se avaliar a eficiência das unidades decisórias analisadas (DMU). A eficiência é uma medida que diferencia essas unidades, permitindo ordená-las e determinar o nível alternativo de insumos que a levaria a produzir em níveis ótimos. Assim, merece ser medida com máxima precisão pois poderá vir a ser utilizada na tomada de decisão, por exemplo na alocação de recursos.

Atualmente existem, dentre outras, duas técnicas para a mensuração de eficiência técnica: uma paramétrica, consistindo na especificação de uma função de produção e uma estrutura de erro composto; e outra não paramétrica, baseada na solução de um problema de programação matemática.

Neste artigo, revisaremos essas alternativas metodológicas. No primeiro caso, abordagem paramétrica, postula-se que a fronteira do conjunto produtivo pode ser representada por uma função de produção caracterizada por parâmetros especificados por métodos estatísticos apropriados. Utilizaremos a especificação de uma forma Cobb-Douglas para a função de produção e estimaremos seus parâmetros através de métodos de simulação estocástica Monte Carlo em Cadeias de Markov (MCMC). Diversos modelos alternativos serão introduzidos pela modificação da parte não negativa da estrutura de erro e, ainda, pelas diferentes hipóteses de hierarquia dos parâmetros. Neste contexto, discutiremos diversos critérios alternativos de seleção de modelos. O segundo método estima não parametricamente a fronteira de produção eficiente. Ao invés de especificar a priori uma particular forma funcional, esta é determinada através de um sistema de equações lineares definidas para calcular os níveis de eficiência referentes a cada observação.

Esses métodos serão aplicados para avaliar as unidades da UFRJ em três anos distintos, 1998, 1999 e 2000. Concatenando as informações existentes, pode-se avaliar, através de análise temporal das unidades, os serviços prestados pelas unidades correspondentes no que se refere ao desempenho educacional da universidade.

Dentre os objetivos deste artigo, destacamos a revisão dos modelos de função de produção estocástica para um único produto e para múltiplos produtos. Os modelos de fronteira estocástica paramétrica para um único produto, introduzidos por Aigner, Lovell & Schmidt (1977), especificam uma função estocástica de um único produto em função dos insumos, nos quais os parâmetros são estimados através de técnicas econométricas tradicionais. A essência

deste método está na especificação do erro estocástico composto. A literatura atual cuida do problema de múltiplos produtos através de três alternativas. A mais simples consiste na construção de um indicador único de produção, obtido pela soma ponderada dos diferentes produtos. Uma segunda alternativa, freqüentemente adotada, é a modelagem de uma função de custo. Para isto é necessário dispor-se de informação de preços de insumos e produtos, o que nem sempre é fácil de se obter. Esta dificuldade pode ser contornada pela introdução dos modelos paramétricos de fator crítico (Banker,1991). Um dos insumos é escolhido, por sua essencialidade, como sendo o fator crítico e será explicado pelos demais insumos e pelos vários produtos, através de uma regressão múltipla com erros compostos. Desta forma esta classe de modelos contorna e estende os modelos econométricos de fronteira de produção evitando as críticas acima mencionadas.

Os modelos de programação matemática, conhecidos como Análise Envoltória de Dados (DEA), cuja referência inicial é de Charnes, Copper & Rhodes (1978) estão baseados, principalmente, na estimativa não-paramétrica de uma fronteira eficiente de produção. A idéia é determinar, de acordo com hipóteses pré-estabelecidas, um sistema de equações lineares que defina o nível de eficiência referente a cada observação. Esses modelos podem ser aplicados tanto a situações de um único produto como ao caso de múltiplos produtos e servirão de referencial na comparação com os modelos que estaremos introduzindo neste artigo.

Este artigo está organizado em quatro seções. A próxima seção está subdividida em três partes. Na primeira, discute-se a fronteira de produção estocástica, detalhando os fundamentos teóricos para mensurar a eficiência e a estrutura de erro composta. Na segunda parte, definem-se as especificações estatísticas dos modelos hierárquicos de fator crítico e suas flexibilidades como alternativa às funções de custo. Por fim, nesta seção, apresenta-se a metodologia de programação matemática DEA. A seção 3 contém a aplicação dos modelos desenvolvidos e os resultados obtidos. Na quarta seção são apresentadas as conclusões, as comparações dos resultados anteriores, além de recomendar extensões e sugestões futuras, que podem servir de subsídios para manter e/ou ampliar a qualidade do ensino superior da universidade.

2. Modelos econométricos de fronteira de produção estocástica

Nesta seção iremos apresentar sumariamente os modelos econométricos de fronteira de produção. Iniciaremos pelos modelos para um único produto e a seguir os modelos de fator

crítico. Ambos serão descritos como modelos hierárquicos e diversas formas alternativas de modelagem do erro composto serão apresentadas. Aspectos da inferência Bayesiana através de MCMC serão também discutidos. Por fim, tem-se as características e conceitos do modelo não paramétrico DEA.

2.1. Modelos de fronteira de produção estocástica

Originalmente este modelo foi proposto por Aigner, Lovell & Schmidt (1977) e Meeusen & Broeck (1977). Algumas revisões importantes desta literatura estão disponíveis nos trabalhos de Forsund, Lovell & Schmidt (1980), Schmidt (1986), Bauer (1990) e Greene (1993).

O modelo tem como principal característica a estimação de uma função de produção em termos de seus fatores de produção. Pode-se obter a produção máxima empregando uma determinada combinação eficiente dos fatores. Em virtude da existência de ineficiências técnicas no uso destes fatores de produção, origina-se o modelo de função de produção estocástica. A grande vantagem dos modelos de fronteira estocástica é que se podem estimar os parâmetros da fronteira e da ineficiência técnica especificando os desvios da fronteira em dois componentes de erros. Um dos componentes de erro é normalmente distribuído com média 0 e variância constante σ_v^2, o qual representa flutuação aleatória que independe das unidades decisórias em estudo. O segundo componente representa a ineficiência técnica dessas unidades. Esta parcela serve para captar a distância entre a fronteira eficiente e os pontos ineficientes e assume distribuições de probabilidades específicas definidas nos reais positivos.

Para fixar idéias e antecipar um pouco nossas aplicações, no caso da universidade, as unidades decisórias serão seus Institutos e os insumos são os recursos disponíveis, quantidades de funcionários e a quantidade de docentes, diferenciados pelo regime de trabalho. Os produtos incluem os alunos formados, as teses e dissertações orientadas, os serviços de extensão realizados e os artigos publicados perseguindo-se a trilogia: ensino, pesquisa e extensão.

Como se pode observar, estamos tipicamente diante de um problema com múltiplos produtos. Para a aplicação dos modelos desta seção teremos de construir um indicador agregado da produção como descreveremos adiante. Assim podemos denotar:

$$Y_i = f(X_i;\beta).\exp(v_i - u_i) \qquad (2.1)$$

onde Y_i é o indicador agregado da produção, X_i um vetor de insumos; β é um vetor de

parâmetros desconhecidos a ser estimado; v_i representa os fatores aleatórios que por hipótese são independentes e identicamente distribuídos com distribuição $N(0;\ \sigma_v^2)$ e independentes de u_i . Este está associado à ineficiência técnica de produção, variáveis aleatórias não-negativas e independentemente distribuídas.

A eficiência técnica de produção para cada observação, pode ser obtida como:

$$e_i = \frac{Y_i}{f(X_i,\beta)\exp(v_i)} = \exp(-u_i)\,. \tag{2.2}$$

Neste artigo a fronteira de eficiência determinística será modelada por uma Cobb-Douglas. Outras especificações da fronteira de produção podem ser encontradas em Medrano (2003), onde além de revisar a literatura existente, exemplos de inferência do ponto de vista Bayesiano são apresentados. A escolha da forma funcional de Cobb-Douglas deve-se ao fato de esta ser comumente utilizada em modelos de fronteira de produção, em razão de suas características simples e de fácil estimação, o que é desejável uma vez que nossos modelos já serão suficientemente complexos nas demais componentes.

A forma funcional da Cobb-Douglas, após a transformação logarítmica e adaptada à nossa aplicação, é dada por:

$$lnY_{it} = \beta_{0it} + \beta_{1it}X_{1it} + \beta_{2it}X_{2it} + \beta_{3it}X_{3it} + v_{it} - u_{it} \tag{2.3}$$

onde $Y_{it}; X_{1it}; X_{2it}$ e X_{3it} são, respectivamente, o indicador de produção, componente principal das variáveis de produção, a quantidade de docentes, a quantidade de funcionários e o volume de recursos orçamentários para despesa de custeio, em reais, do i-ésimo instituto no período t. Com exceção do intercepto, os valores de β's podem ser interpretados como produtividades parciais dos recursos e da mão de obra.

Podemos rescrever os modelos acima como modelos hierárquicos fazendo:

$$(Y_i \,|\, \beta, \sigma_v^2, u_i) \approx N\left[f(X_i,\beta) - u_i, \sigma_v^2\right]$$

$$ui \approx F[\mu, \sigma_u^2]$$

onde u_i pode ser modelado por uma função de distribuição sobre os reais positivos.

Tipicamente elas são membros de uma das famílias: (i) $u_i \approx N\left|\mu, \sigma_u^2\right| I_{(0,\infty)}$, $\sigma_u^2 > 0$, normal truncada, (ii) $u_i \approx Ga(m, \sigma_u)$, $m, \sigma_u > 0$, gama com parâmetros reais positivos desconhecidos, ou (iv) $u_i \approx LN(\mu, \sigma_u^2)$ log-normal. Esta última mesmo supondo-se $\mu = 0$ tem a moda fora da origem o que é parcimonioso e altamente desejável. Alternativamente, podemos introduzir covariáveis para o segundo nível do modelo com o objetivo de explicar a diferença de eficiências das unidades. Isto será feito mais abaixo no contexto dos modelos de fator crítico.

A abordagem Bayesiana foi utilizada para se fazerem inferências nesses modelos hierárquicos de função de produção estocástica, usando-se MCMC –Monte Carlo Markov Chain (Gilks et al, 1995). Ao utilizar o método MCMC, foi necessário especificar *prioris* para os parâmetros $\beta, \sigma_v^2, \sigma_u^2$ e μ dos modelos:

i) $\beta \approx \mathrm{Np(b, B)}$ e $\mu \approx \mathrm{N}(a, \mathrm{A})$

ii) $\sigma_v^2 \approx Ga(c/2, C/2)$ e $\sigma_u^2 \approx Ga(d/2, D/2)$

onde a,b,c,d,A,B,C,D são os hiperparâmetros do modelo. Estes são positivos e escolhidos de tal forma que a distribuição *a priori* seja vaga, porém, própria. Ao aplicar MCMC são obtidas as distribuições posteriores condicionais completas, onde os teoremas no Anexo (1) resumem os principais resultados.

2.2. Modelos de fator crítico

O modelo de fator crítico foi introduzido por Banker (1991) e é uma extensão das técnicas de estimação de fronteiras de produção de Aigner (1977) ao caso de múltiplos produtos. O princípio fundamental desta modelagem consiste em formular modelos em termos de um fator crítico, aquele insumo mais essencial, identificando as combinações de produtos, além dos demais fatores que o explicam, permitindo assim avaliar a eficiência. Intuitivamente, estaríamos medindo a eficiência pelo excedente entre o valor observado do fator e o predito a partir dos produtos, isto é, quanto usamos deste fator em excesso ao que seria esperado de sua utilização.

Estes modelos serão, também, formalizados como modelos hierárquicos Bayesianos ou modelos multiníveis, como às vezes é chamado na literatura clássica (Bryk & Raudenbush, 2002). A eventual estrutura hierárquica das unidades decisórias é incorporada na modelagem. Por exemplo, no caso da universidade, é bem sabido que a grande variabilidade na produção dos institutos pode, em parte, ser explicada pela estrutura dos centros. Aqueles pertencentes a um

mesmo centro tendem a ter comportamento similar. Na modelagem hierárquica a variabilidade total dos dados é parcimoniosamente decomposta segundo os níveis postulados pela hierarquia. De outra forma, poderíamos dizer que o componente de erro que descreve os dados é decomposto segundo uma particular estrutura hierárquica imposta sobre os parâmetros, possibilitando explicar grande parte da variabilidade dos dados.

Os modelos hierárquicos descritos a seguir têm, portanto, dois níveis distintos decorrentes dos diferentes institutos que possuem perfis semelhantes aos seus respectivos centros. A primeira equação representa o primeiro nível do modelo. A equação básica deste nível relaciona a força de trabalho do docente, fator crítico ou principal insumo, que pode ser explicada por alguns resultados finais de produção, tais como: aulas ministradas, alunos orientados e trabalhos publicados. A segunda equação do modelo retrata o segundo nível do problema, estabelece a relação entre os coeficientes técnicos da produtividade do fator crítico com a disponibilidade de outros fatores, bem como a similaridade destes fatores nas unidades de um mesmo centro.

A estimação dos parâmetros dos modelos é realizada por meio de inferência Bayesiana e de simulação de Monte Carlo em Cadeias de Markov - MCMC, em que se admitem densidades a priori não-informativas para cada um dos hiperparâmetros. A análise a posteriori é efetuada por meio de algoritmos de simulação MCMC, aqui adotado o algoritmo de Amostrador de Gibbs (Casella & George, 1992).

Três variantes de modelos hierárquicos de fator crítico são propostos neste artigo, os quais diferem segundo a hipótese de permutabilidade utilizada na descrição dos parâmetros do segundo nível. No primeiro modelo, somente o intercepto pode variar entre as unidades pesquisadas segundo o Centro a que pertence. No segundo modelo, mais geral, todos os coeficientes das covariáveis podem variar de acordo com o Centro. E, finalmente, no terceiro modelo, a variação dos coeficientes técnicos (regressores) é explicada por fatores de produção ou insumos mensurados a nível de cada centro.

Assim, segue-se a especificação dos três modelos de fator crítico considerados. Denota-se por i as unidades institucionais, j, os centros da universidade, t, refere-se ao tempo, k, índice de variável Y, e l, índice de variável Z. X é o vetor de produtos, Y e Z são os insumos. Onde Y é o insumo principal (fator crítico) e Z indicado pelo número de funcionários e o volume de recursos orçamentários. Observe que este método permite utilizar múltiplos insumos e produtos.

$$\text{Modelo 1: } Y_{ijt} = \beta_{0jt} + \sum_{k=1}^{3} \beta_{kj} X_{ijkt} + \varepsilon_{ijt}\,, \tag{2.4}$$

$$\beta_{0jt} = \alpha_{0t} + w_{0jt}$$

$$\text{Modelo 2: } Y_{ijt} = \beta_{0jt} + \sum_{k=1}^{3} \beta_{kj} X_{ijkt} + \varepsilon_{ijt}\,, \tag{2.5}$$

$$\beta_{0jt} = \alpha_{0t} + w_{0jt}$$

$$\beta_{kj} = \alpha_{k} + w_{kj}$$

$$\text{Modelo 3: } Y_{ijt} = \beta_{0jt} + \sum_{k=1}^{3} \beta_{kj} X_{ijkt} + \varepsilon_{ijt}\,, \tag{2.6}$$

$$\beta_{0jt} = \alpha_{0t} + w_{0jt}$$

$$\beta_{kj} = \alpha_{0k} + \sum_{l=1}^{2} \alpha_{kl} Z_{jlt} + w_{kj}\,.$$

Vale ressaltar, que os modelos acima admitem que a produção por docente varie com o tempo.

Os resíduos ajustados são simplesmente a diferença entre os valores do volume de fator crítico e os valores ajustados para este volume pelo modelo. As diferenças positivas indicam que os docentes gastam mais carga horária do que foi estimada para a produção do instituto. Ao contrário, diferenças negativas refletem uma eficiência produtiva do instituto considerado.

Uma outra maneira de identificar os institutos eficientes seria reestruturando a primeira equação dos modelos de fator crítico. Observe que esta expressão, comum aos três modelos, pode também ser especificada da seguinte forma:

$$Y_{ijt} = \beta_{0jt} + \sum_{k=1}^{3} \beta_{kj} X_{ijkt} + \varepsilon^{*}_{ijt} - u^{*}_{ijt} \tag{2.7}$$

onde $\varepsilon_{ijt} = \varepsilon^{*}_{ijt} - u^{*}_{ijt}$.

A produção máxima por docente que o instituto i pode ter é determinada pela função $f(X_{ijkt};\beta)$, onde a eficiência pode ser denotada por $e_{it} = \exp(-u^{*}_{ijt})$.

Esta especificação indica que, quanto maior o valor da eficiência (e_{it}), menor o denominador, ou seja, menos produtivo pode ser considerado o instituto i no período t. Assim, ao

contrário dos modelos de fronteira de produção estocástica, para os modelos de fator crítico admite-se que quanto menores os valores da eficiência, melhor o instituto.

2.3 Modelos DEA

A mensuração dos níveis de eficiência pode ser feita mediante o uso da metodologia conhecida como Análise Envoltória de Dados - DEA, cuja referência inicial é de Charnes, Cooper & Rhodes (1978), ou simplesmente modelo CCR. O modelo CCR e o modelo BCC (Banker, Charnes e Cooper,1984) formam o núcleo das estruturas analíticas que constituem o método DEA. Este método trata de uma abordagem não-paramétrica, em que os modelos se baseiam em programação matemática na construção da fronteira eficiente de produção a partir do conjunto de unidades tomadoras de decisão (DMU's), que podem ser definidas como firmas, departamentos, divisão ou unidades administrativas.

O método DEA otimiza cada observação individual como critério de determinação da fronteira linear por partes, que compreende o conjunto de DMU's Pareto Efciente[1].

Tecnicamente, o modelo DEA utiliza a otimização de programação matemática linear para construir uma fronteira de produção empírica, ou superfície envoltória de máximo desempenho. Isto permite que se identifiquem unidades de referência, cujos índices de desempenho servem como referencial para as demais unidades, posicionadas sob a superfície envoltória.

A fronteira utilizada no presente estudo é proveniente do modelo BCC, o qual pode ser descrito da seguinte forma:

max h; sujeito a

$$
\begin{aligned}
-\sum_{k=1}^{n} O_{yk}\lambda_K + hO_{yo} &\le 0 \qquad \text{para} \quad y=(1,...,s) \\
\sum_{k=1}^{n} I_{xk}\lambda_K &\le I_{xo} \qquad \text{para} \quad x=(1,...,m) \\
\sum_{k=1}^{n} \lambda_K &= 1 \\
\lambda_k &\ge 0 \qquad \text{para} \quad k=(1,...,n)
\end{aligned}
\tag{2.8}
$$

[1] Nenhum dos produtos (resp. insumo) podem ser aumentados sem que algum outro (resp.insumo) seja reduzido ou algum produto (resp. insumo), aumentado.

onde h é a eficiência associada a unidade; m é o número total de insumos; s é o número total de produtos; n é o número total de unidades; I é quantidade de insumo X para a unidade K; O é a quantidade de produto Y para unidade K.

Este problema é resolvido para cada DMU, de modo que existam n problemas de programação matemática a serem resolvidos. A terceira restrição, de convexidade, possibilita a presença de retornos variáveis de escala.

Os modelos DEA comparam a medida virtual de produtos com a medida virtual de insumos, com pesos escolhidos de modo que cada DMU seja representada da forma mais eficiente possível. Esta característica deve ser consistente com os dados e com a restrição de não ultrapassar a fronteira de produção.

3. Análise dos resultados

Os resultados obtidos após o ajuste dos modelos especificados anteriormente aos dados reais da UFRJ referentes aos anos de 1998, 1999 e 2000 são analisados nesta seção. A análise inclui somente os 37 institutos e os 6 centros que têm sob sua responsabilidade as atividades de ensino, pesquisa e extensão. Três grupos de centros de ensino foram formados: o primeiro grupo (Centro 1) é composto pelos Centros de Tecnologia - CT e Ciências Matemáticas e da Natureza - -CCMN, o segundo grupo (Centro 2) é formado pelos Centros de Filosofa e Ciências Humanas – CFCH, de Ciências Jurídicas e Econômicas – CCJE e de Letras e Artes – CLA, e, por último, o Centro 3, tem-se o Centro de Ciências da Saúde – CCS.

A aplicação envolve a construção de medidas dos principais produtos da universidade: ensino, pesquisa e extensão, e levam em consideração os insumos usuais: capital e trabalho. Neste contexto, alguns indicadores de qualidade educacional foram construídos e classificados em três categorias:

- índice de publicação: construído a partir da soma ponderada do número de artigos publicados em periódicos internacionais e nacionais, o número de congressos nacionais e internacionais, e o número de outras publicações.
- índice de orientação: calculado através da soma ponderada das dissertações de mestrado e do número de teses de doutorado orientadas por professores do instituto no período.
- quantidade de alunos: esta variável foi decomposta em duas, meramente por dificuldades operacionais: a primeira, trata-se dos alunos inscritos nas disciplinas da graduação, e a segunda, contém os alunos matriculados nos programas de pós-graduação.

Para aplicar o modelo de fronteira de produção estocástica paramétrico, houve a necessidade de se construir um único indicador que representasse estas dimensões. Optou-se por utilizar a análise fatorial a partir do método de componentes principais, que consiste em estimar os parâmetros do modelo fatorial pré-estabelecendo algumas hipóteses (Jonhson, 1999). Seus objetivos principais são reduzir a dimensionalidade dos dados e facilitar sua interpretação, onde o primeiro fator contabiliza a maior parcela de variação dos dados, o segundo contabiliza a segunda maior parcela e assim por diante. Sucessivos fatores explicam parcelas cada vez menores da variância total e são independentes uns dos outros. Através da extração por componentes principais o primeiro fator explica, aproximadamente, 60%, 63% e 61% da variabilidade total das variáveis em estudo nos anos de 1998, 1999 e 2000, respectivamente. Sendo assim, este indicador representa significativamente as três dimensões dos principais produtos da universidade.

No caso dos indicadores de insumos, trabalhou-se com três agregações:

- quantidade dos professores: foram utilizadas como diferencial as cargas horárias disponíveis na universidade, assim como o grau de escolaridade dos docentes por unidade da UFRJ. Foi ponderado o número de professores no regime de 20 horas semanais de trabalho, 40 horas semanais e aqueles de dedicação exclusiva, respeitando o nível de escolaridade de cada um, aqueles que são graduados e os que são titulados como mestres e doutores.
- número de funcionários: quantidade de funcionários técnicos e administrativos para cada unidade de estudo.
- volume de recursos orçamentários: disponibilidade de recursos de custeio, em reais, atribuídos a cada unidade.

Definir pesos com o compromisso de refletir a percepção e a importância relativa de cada variável a cada unidade é uma tarefa difícil e questionável. Assim, a obtenção dos pesos atribuídos às variáveis seguem as instruções gerais da Gratificação de Estímulo à Docência - GED, utilizando cada item de atividade de ensino, orientações e avaliações das produções e das extensões. Os quadros com os ponderadores são apresentados no Anexo 2.

Os modelos utilizados são classificados em: modelos de fator crítico; modelos de função de produção estocástica e os modelos DEA. Assim, a análise dos resultados foi composta por algumas etapas relacionadas a cada método proposto. Primeiramente, ajustam-se e interpretam-se os três modelos de fator crítico propostos na seção 2.2 para três anos distintos (1998, 1999 e 2000). Calculam-se medidas de diagnósticos dos respectivos modelos, com o objetivo de selecionar o mais adequado. O modelo escolhido será então analisado e confrontado com as

outras metodologias. Posteriormente, nos modelos paramétricos, estes três anos contribuíram para inclusão do efeito temporal nas unidades. Por fim, a análise apresenta, para fins de comparação, a correlação de Spearman obtida por cada modelo.

Modelos de fator crítico

As primeiras análises são decorrentes dos três modelos hierárquicos de fator crítico descritos na seção anterior. Um destes três modelos será escolhido e representará a metodologia de Fator Crítico sendo, assim, comparado às outras metodologias abordadas. Os critérios de seleção do modelo utilizados foram o da verossimilhança preditiva (VP)[2] e o da minimização do desvio preditivo esperado (EPD), ver Anexo 3 para maiores detalhes. A Tabela 3.1 mostra de forma suscinta, os valores destas medidas para os três modelos ajustados para os três anos considerados.

Tabela 3.1 - Medidas de diagnósticos

Anos	Modelo 1		Modelo 2		Modelo 3	
	EPD	VP	EPD	VP	EPD	VP
1998	12.50	-19.45	6.79	-7.95	**6.74**	**-7.78**
~~1999~~	~~9.27~~	~~13.89~~	~~7.52~~	~~-9.87~~	~~**7.37**~~	~~**-9.49**~~
2000	14.16	-21.75	9.09	-13.39	**8.79**	**-12.77**

Analisando-se estas medidas de diagnósticos, pode-se perceber que o modelo 3 é o mais adequado para os três anos, pois possui o menor desvio preditivo esperado e a maior verossimilhança preditiva. Este modelo considera que, tanto o intercepto como os coeficientes de produtividade relacionados aos docentes, são explicados por outras variáveis, tais como os recursos e os funcionários, dependendo do centro ao qual pertencem. Assim, para evitar uma proliferação de modelos a serem comparados com as outras metodologias, o modelo 3 foi o escolhido para ser analisado e confrontado com as outras abordagens.

A Tabela 3.2 apresenta as estimativas de cada coeficiente do primeiro nível nos três anos distintos do modelo 3 de Fator Crítico (FC), já incorporando as duas componentes de erro.

---

[2] O Logaritmo da Verossimilhança Preditiva (VP) permite avaliar e comparar modelos através de suas densidades preditivas, medindo o grau de ajustamento do modelo aos dados. $VP = \log[2\pi - 1/2\sum_{i=1}^{n} f(x_i, \beta) - 1/2\sum_{i=1}^{n} e_i^2]$, denota-se e_i^2 o erro quadrático médio.

Verifica-se que alguns coeficientes obtiveram resultados não-signifcativos, pois possuem intervalos de credibilidade incluindo o zero, principalmente as estimativas dos coeficientes das variáveis alunos inscritos nas disciplinas de graduação e índice de orientações. Apesar de não significativas, é possível que estas variáveis sejam importantes para análise da produtividade do instituto. Desta forma, preferiu-se não retirá-las do modelo. O coeficiente que apresentou maior valor foi o índice de publicações, ou seja, o impacto deste índice no fator crítico é o maior na maioria dos anos e dos centros analisados.

Tabela 3.2 - Parâmetros estimados do modelo de Fator Crítico - FC

Anos	1998		1999		2000	
	Média	Desvio	Média	Desvio	Média	Desvio
Centro 1: CT e CCMN						
Intercepto	1.07	0.11	1.03	0.11	1.19	0.12
Publicações	0.99	0.21	0.89	0.17	0.81	0.21
Alunos Grad	0.02	0.13	-0.01	0.16	0.16	0.17
Orientações	0.09	0.16	0.07	0.22	-0.01	0.22
Alunos Pós-Grad	0.47	0.11	0.35	0.08	0.57	0.16
Centro 2: CFCH, CCJE e CLA						
Intercepto	1.48	0.15	1.50	0.17	1.68	0.26
Publicações	1.40	0.31	1.18	0.35	1.67	0.64
Alunos Grad	0.13	0.26	-0.14	0.50	0.42	0.37
Orientações	0.26	0.47	0.59	0.53	0.12	0.34
Alunos Pós-Grad	0.06	0.09	0.06	0.11	0.05	0.10
Centro 3: CCS						
Intercepto	0.98	0.15	0.86	0.19	0.93	0.26
Publicações	0.53	0.15	0.90	0.29	0.54	0.27
Alunos Grad	-0.10	0.40	-0.05	0.67	-0.26	0.98
Orientações	0.12	0.35	-0.24	0.51	0.30	0.53
Alunos Pós-Grad	0.55	0.20	0.41	0.18	0.49	0.21

O modelo de fator crítico admite que a produção por docente varie com o tempo. Isto foi feito através da variação temporal das variáveis, dos níveis de eficiência e do intercepto ($\beta_{0jt} = \alpha_{0t} + \varepsilon_{jt}$) onde, t são os anos e j os centros). A Tabela 3.3 apresenta as estimativas de

cada coeficiente do primeiro nível para o modelo de fator crítico temporal. Novamente, o coeficiente de índice de publicação apresentou o maior impacto no fator crítico para os três conjuntos de centros analisados.

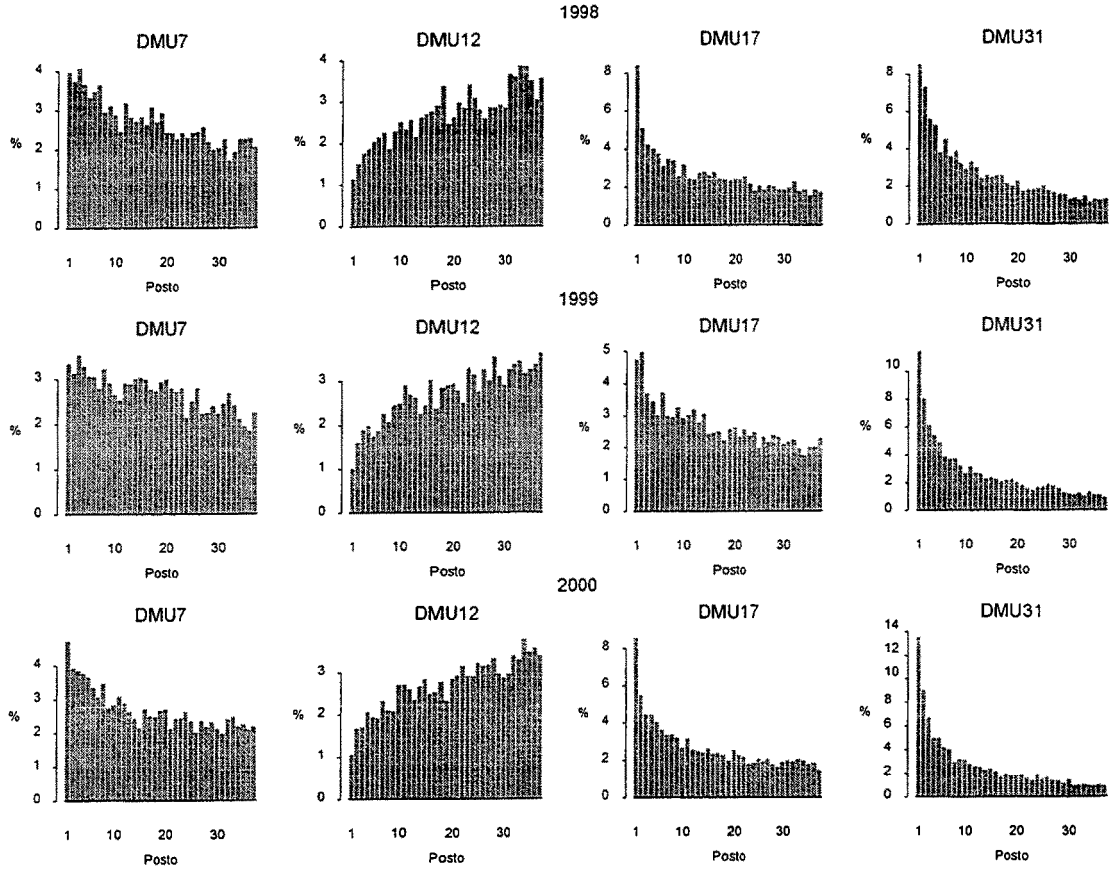
Tabela 3.3 - Parâmetros estimados do modelo de fator crítico temporal - FCT

Anos	1998		1999		2000	
	Média	Desvio	Média	Desvio	Média	Desvio
Centro 1: CT e CCMN						
Intercepto	1.02	0.41	0.93	0.43	1.07	0.41
Publicações	0.76	0.40	0.76	0.40	0.76	0.40
Alunos Grad	0.04	0.32	0.04	0.32	0.04	0.32
Orientações	0.15	0.40	0.15	0.40	0.15	0.40
Alunos Pós-Grad	0.46	0.25	0.46	0.25	0.46	0.25
Centro 2: CFCH, CCJE e CLA						
Intercepto	1.08	0.39	1.11	0.41	1.10	0.43
Publicações	0.64	0.60	0.64	0.60	0.64	0.60
Alunos Grad	0.34	0.58	0.34	0.58	0.34	0.58
Orientações	0.39	0.61	0.39	0.61	0.39	0.61
Alunos Pós-Grad	0.15	0.21	0.15	0.21	0.15	0.21
Centro 3: CCS						
Intercepto	0.87	0.41	0.88	0.40	0.84	0.40
Publicações	0.53	0.37	0.53	0.37	0.53	0.37
Alunos Grad	0.03	0.81	0.03	0.81	0.03	0.81
Orientações	0.20	0.70	0.20	0.70	0.20	0.70
Alunos Pós-Grad	0.43	0.38	0.43	0.38	0.43	0.38

Uma facilidade decorrente da inferência Bayesiana via MCMC é a possibilidade de se obter a distribuição de quantidades de interesse derivadas dos parâmetros originais do modelo. Nesta categoria inclui-se a distribuição dos postos relativos à medida de eficiência. A Figura 1 evidencia as estimativas da distribuição dos postos associados aos níveis de eficiência das unidades para todas as interações do modelo de fator crítico temporal. Os níveis de eficiência foram ordenados crescentemente e obtidos pela fórmula $e_{it} = \exp(-u_{it})$. Nos modelos de fator crítico, ao contrário dos modelos de fronteira de produção estocástica e do DEA, quanto menores os valores da eficiência (e_{it}), mais produtivo pode ser considerado o instituto. Destacaram-se algumas unidades para serem observadas nos três anos estudados. Nota-se que os Institutos de Biofísica,

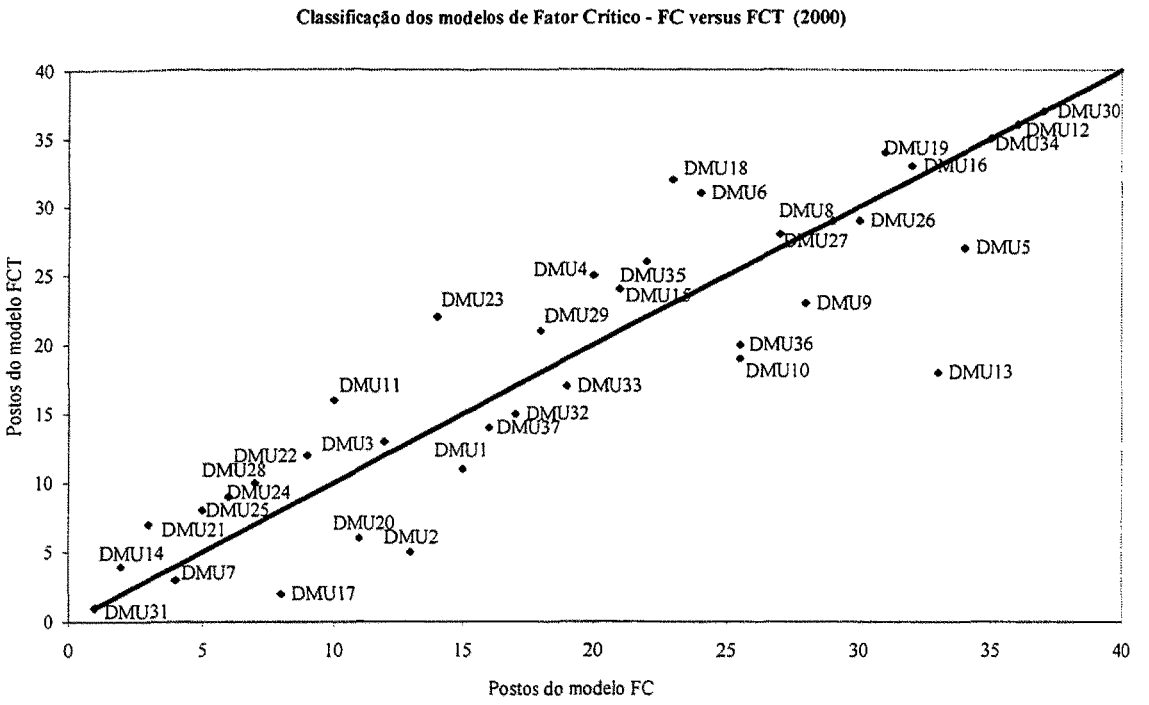
de Macromoléculas e a Faculdade Nacional de Direito, possuem as maiores freqüências nos menores postos de eficiência, o que indica maior desempenho nas suas produtividades. Em contraposição, as unidades pertencentes ao Centro de Letras e Artes apresentam maiores freqüências nos maiores postos de eficiência, o que as revelam com baixa produtividade, para a maioria das interações e em todos os anos.

Figura 1 - Distribuição dos postos associados aos níveis de eficiências do Modelo Fator Crítico ao longo dos três anos para algumas unidades selecionadas: DMU7(Instituto de Macromoléculas), DMU12(Escola de Música), DMU17(Faculdade Nacional de Direito), DMU31(Instituto Biofísica)



Com o objetivo de verificar se há semelhança na classificação das unidades do modelo de fator crítico atemporal e temporal, a Figura 2 apresenta a ordenação crescente das médias dos níveis de eficiência de ambos os modelos para o ano de 2000 pesquisado. No eixo X está a ordenação das unidades para o modelo de fator crítico temporal, e no eixo Y, a ordenação das unidades do modelo de fator crítico temporal. Observa-se que quanto mais próximo da reta estiverem os pontos, há coincidência entre os resultados dos modelos.

Figura 2 - Relação entre os postos dos modelos de fator crítico - 2000



Assim, dirige-se a atenção aos Institutos de Biofísica, Macromoléculas, COPPEAD, entre outros, que possuem maiores destaques na produtividade, enquanto os Institutos dos Centros CFCH e CLA possuem baixos níveis de produção.

Modelos de fronteira de produção estocástica

Nesta etapa analisaram-se os modelos de fronteira de produção estocástica, com forma funcional tipo Cobb-Douglas e eficiência técnica, a ser estimada, com distribuição lognormal. Nestes modelos, duas abordagens de fronteira de produção são apresentadas: na primeira, a equação é definida para os institutos da universidade; na segunda, o modelo incorporou a estrutura hierárquica da universidade, a qual relaciona os institutos aos seus respectivos centros.

Os valores estimados dos parâmetros relacionados aos fatores de produção dos institutos da UFRJ para os diferentes anos pesquisados do modelo de fronteira de produção estocástica são apresentados na Tabela 3.4.

Tabela 3.4 - Parâmetros estimados do modelo de fronteira de produção estocástica – FPE

Coeficientes	1998		1999		2000	
	Média	Desvio	Média	Desvio	Média	Desvio
Intercepto	1.70	0.10	1.78	0.08	1.74	0.08
Docentes	0.32	0.24	0.26	0.19	0.09	0.21
Funcionários	0.53	0.24	0.30	0.20	0.45	0.21
Recursos	-	-	0.50	0.10	0.48	0.10
γ	0.10	0.09	0.27	0.21	0.27	0.21
σ_{υ}^2	0.04	0.04	0.06	0.05	0.05	0.05
σ_{w}^2	0.34	0.10	0.15	0.06	0.15	0.06

Com exceção do intercepto, os valores dos coeficientes podem ser interpretados como produtividades parciais de recurso e da mão de obra. Dentre as principais análises, sobressai a comprovação de que quando um acréscimo proporcional nos insumos resultaria num acréscimo maior que o proporcional nos produtos, visto que a soma dos coeficientes tem uma distribuição que pode ser descrita por: média = 0.85, mediana = 0.85, 1º quartil = 0.64 e 3º quartil = 1.06 no ano de 1998. Para o ano de 1999 têm-se: média = 1.06, mediana = 1.06, 1º quartil = 0.88 e 3º quartil = 1.23. Considerando o ano de 2000 têm-se: média = 1.02, mediana = 1.02, 1º quartil = 0.85 e 3º quartil = 1.18. Ou seja, em todos os anos pesquisados as somas dos coeficientes médios são superiores a 1. Percebe-se também que o coeficiente de maior valor está relacionado à variável de recursos, ou seja, os institutos considerados mais produtivos dispõem de mais capital. Os resultados da referida verificação apontam o coeficiente recursos como significativo, ou seja, o intervalo de credibilidade não contém o zero.

Analisando-se as eficiências técnicas estimadas, em média, os níveis de eficiência estão em torno de 0.37. Os institutos que apresentam piores desempenhos, quando comparados com a média, são aqueles pertencentes principalmente ao CLA. Dentre os que possuem melhor capacidade de absorção da produtividade, têm-se a COPPE, o Instituto de Macromoléculas, a COPEAD e o Instituto de Biofísica. Algumas unidades apresentam trajetória crescente entre os anos, tais como a faculdade de farmácia e de medicina.

A segunda abordagem do modelo de fronteira de produção estocástica admitiu a estrutura hierárquica da universidade. Na Tabela 3.5 apresentam-se as estimativas dos parâmetros para este modelo.

Tabela 3.5 - Parâmetros estimados do modelo de fronteira de produção estocástica hierárquica – FPEH

Anos	1998		1999		2000	
	Média	Desvio	Média	Desvio	Média	Desvio
Centro 1: CT e CCMN						
Intercepto	1.83	0.18	1.81	0.11	1.67	0.11
Docentes	-0.31	0.40	0.35	0.20	0.08	0.21
Funcionários	1.52	0.43	0.02	0.25	0.19	0.26
Recursos	-	-	0.80	0.11	0.79	0.12
Centro 2: CFCH, CCJE e CLA						
Intercepto	1.56	0.15	1.58	0.09	1.56	0.09
Docentes	0.26	0.35	0.51	0.20	0.34	0.23
Funcionários	0.24	0.48	0.00	0.23	0.17	0.26
Recursos	-	-	-0.06	0.22	0.07	0.24
Centro 3: CCS						
Intercepto	1.71	0.13	1.78	0.09	1.71	0.11
Docentes	0.38	0.36	0.59	0.20	0.44	0.22
Funcionários	0.29	0.30	0.05	0.17	0.23	0.20
Recursos	-	-	0.04	0.11	-0.14	0.21

Observa-se que para o CT e de CCMN o coeficiente de maior valor é correspondente a variável de recursos, ou seja, o capital investido nos institutos deste centro são os que mais contribuem para o aumento da produtividade. Já para os centros restantes, a carga horária dos docentes apresenta valores mais elevados. Neste caso, a dedicação dos docentes é de maior importância na produtividade dos institutos pertencentes a estes centros.

A Tabela 3.6 mostra os resultados das medidas de diagnóstico de ambos os modelos analisados ano a ano. Nota-se que o modelo de Fronteira Estocástico, quando considera a estrutura hierárquica dos dados, apresenta melhor ajuste. Este possui o menor desvio preditivo esperado e a maior verossimilhança preditiva. Assim, foi este o modelo escolhido para incorporar a variação temporal nos níveis de eficiência.

Tabela 3.6 - Medidas de diagnósticos fronteira de produção estocástica

Anos	FPE		FPEH	
	EPD	VP	EPD	VP
1998	24.29	-31.72	13.40	-19.13
1999	10.79	-15.52	4.83	-1.19
2000	11.08	-16.10	5.84	-4.86

A partir de agora passaremos a fazer uma análise intertemporal no modelo de fronteira de produção. A componente temporal admite variação na eficiência e no intercepto ($\beta_{0jt} = \alpha_{0t} + \varepsilon_{jt}$ onde, t são os anos e j os centros).

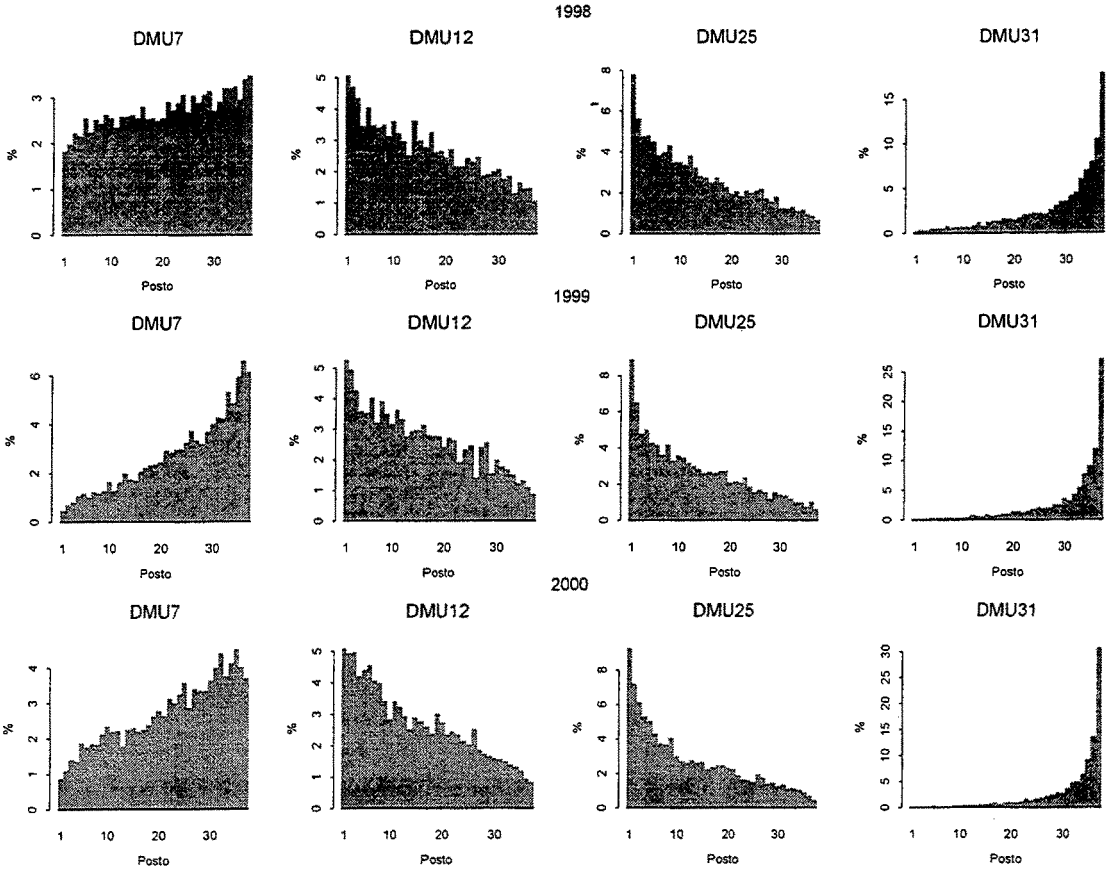
A Tabela 3.7 apresenta os parâmetros estimados do modelo de fronteira de produção estocástica hierárquica temporal. Os resultados mostram que a contribuição dos docentes e dos recursos disponíveis tem impacto importante sobre as estimativas da eficiência relativa dos institutos. Observa-se, para os Centros CT e CCM que, quando ocorre um acréscimo proporcional nos insumos, resultaria num acréscimo mais que proporcional na produção, pois os coeficientes têm uma distribuição descrita por: média = 1.1, mediana = 1.1, 1º quartil = 1.0, 3º quartil = 1.2.

Tabela 3.7 - Parâmetros estimados do modelo de fronteira de produção estocástica hierárquica temporal - FPEHT

Anos	1998		1999		2000	
	Média	Desvio	Média	Desvio	Média	Desvio
Centro 1: CT e CCMN						
Intercepto	1.68	0.18	1.67	0.13	1.74	0.13
Docentes	0.12	0.15	0.12	0.15	0.12	0.15
Funcionários	0.23	0.20	0.23	0.20	0.23	0.20
Recursos	0.76	0.08	0.76	0.08	0.76	0.08
Centro 2: CFCH, CCJE e CLA						
Intercepto	1.53	0.08	1.59	0.07	1.55	0.08
Docentes	0.48	0.12	0.48	0.12	0.48	0.12
Funcionários	0.01	0.15	0.01	0.15	0.01	0.15
Recursos	0.39	0.04	0.39	0.04	0.39	0.04
Centro 3: CCS						
Intercepto	1.72	0.09	1.76	0.07	1.77	0.07
Docentes	0.63	0.12	0.63	0.12	0.63	0.12
Funcionários	0.07	0.10	0.07	0.10	0.07	0.10
Recursos	0.37	0.04	0.37	0.04	0.37	0.04

Na Figura 3 observa-se a distribuição dos postos associados aos níveis de eficiência para todas as interações MCMC do modelo de fronteira de produção temporal. Como pode ser visto, os Institutos de Macromoléculas e Biofísica possuem as maiores freqüências nos postos mais altos de eficiência, indicando-os como eficientes. Enquanto que as unidades do CLA apresentam menores freqüências, o que determina sua baixa produtividade.

Figura 3 - Distribuição dos postos associados aos níveis de eficiências do Modelo de Fronteira de Produção Estocástica ao longo dos três anos para algumas unidades selecionadas: DMU7 (Instituto de Macromoléculas), DMU12(Escola de Música), DMU25 (Escola de Educação Física e Desportos), DMU31(Instituto Biofísica)



Modelo DEA

A terceira análise utiliza os modelos DEA. Estes consistem de um sistema de equações lineares e tem por objetivo maximizar a produção dos institutos sem utilizar um número adicional de docentes, funcionários e recursos. Após uma análise gráfica relacionando insumos com produtos, verificou-se que estas relações se adaptam melhor à fronteira VRS calculada pelo modelo BCC, o que permite rendimentos variáveis de escala.

Para esta análise foram propostos dois modelos: o primeiro utiliza o DEA com múltiplos produtos (DEAM) e o segundo com um único produto (DEA1), que é obtido através da mesma componente principal empregada para o modelo de FPE.

Dentre os resultados que a modelagem DEA fornece incluem-se: o escore de eficiência das unidades, a freqüência com a qual uma unidade eficiente aparece como referência a uma outra

unidade dada como ineficiente e o conjunto de referência de unidades eficientes para as ineficientes.

Analisando primeiramente o modelo com um único produto (DEA1), quatro institutos foram considerados eficientes nos três anos pesquisados: a COPPE, o Instituto de Macromoléculas, Observatório do Valongo e o Instituto de Biofísica. Cada qual pertence ao hiperplano de referência μ para as unidades ineficientes. A unidade eficiente considerada com maior número de referência foi o Instituto de Biofísica com freqüência de 32, 27 e 29 para os anos de 1998, 1999 e 2000, respectivamente. Estas freqüências revelam a importância relativa atribuída ao desempenho do Instituto de Biofísica, sendo utilizado como referência por outras 32 unidades no ano de 1998. Porém, uma unidade ineficiente pode extrair esta importância de mais de uma unidade eficiente, o que compõe o seu conjunto de referência.

Da comparação da DMU ineficiente com o seu conjunto de referência, podem-se extrair estimativas dos insumos e produtos necessários para que a referida DMU atinja sua eficiência. A Tabela 3.8 exemplifica estes resultados para algumas unidades. Por exemplo, no ano de 1999, a Faculdade de Administração e Ciências Contáveis poderia atingir a fronteira de eficiência aumentando o índice de publicações de 0.2 para 0.9, reduzindo a equipe de funcionários de 56 para 41 e mantendo o seu corpo docente e os seus recursos disponíveis.

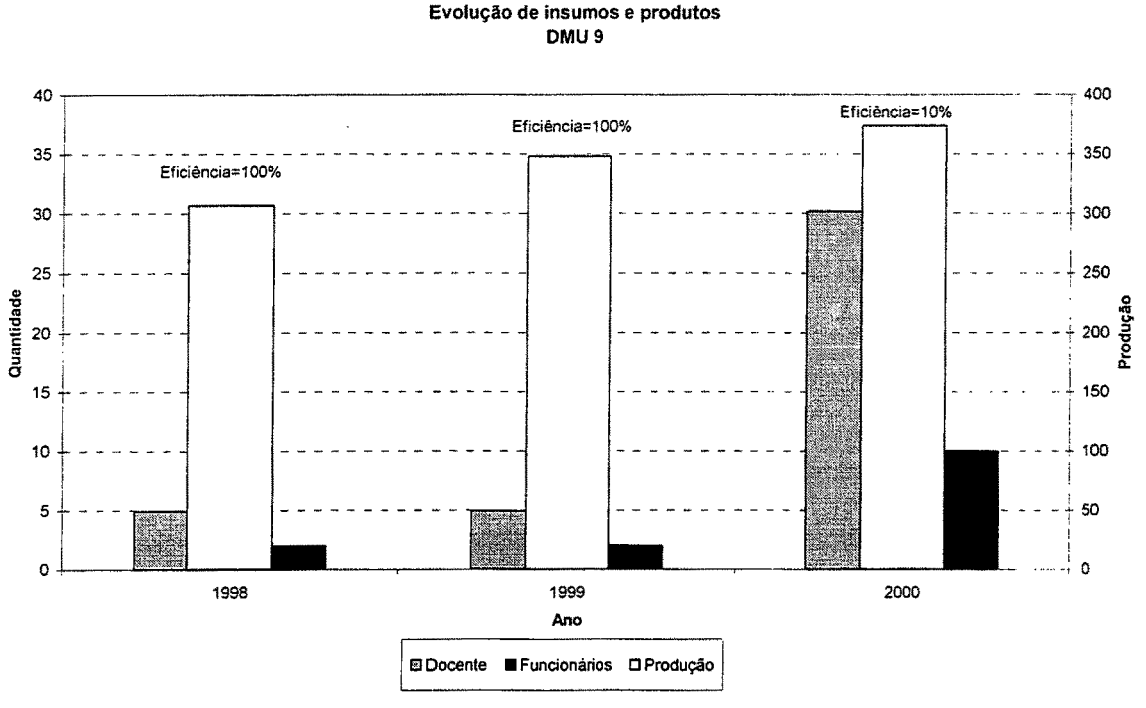
Tabela 3.8 - Estimativas nos inputs e outputs de melhoras para as unidades ineficientes - Modelo DEA-1

Anos	Valor Original				Valor Projetado			
	PUB	DOC	FUNC	REC	PUB	DOC	FUNC	REC
			1998					
Faculdade de Administração e Ciências Contáveis	0.0	83	56	-	1.5	83	54	-
Escola de Educação Física e Desportos	0.0	104	120	-	1.7	104	67	-
			1999					
Faculdade de Administração e Ciências Contáveis	0.2	88	56	140 955	0.9	88	41	140 955
Escola de Educação Física e Desportos	0.1	113	117	435 505	1.7	113	74	435 505
			2000					
Faculdade de Administração e Ciências Contáveis	0.1	91	58	114 963	0.7	91	49	114 963
Escola de Educação Física e Desportos	0.1	118	118	642 084	1.9	118	74	596 050

Ao aplicar o modelo com múltiplos produtos (DEAM), foi observado um aumento do número de unidades pertencentes à fronteira. Sete institutos se apresentam eficientes durante os 3 anos estudados, são eles: a COPPE, Instituto Biofísica, a Escola de Engenharia, o Instituto de Macromoléculas, dentre outros.

Um caso interessante que merece ser comentado é o ocorrido com a unidade denominada Observatório do Valongo - OV. No ano de 2000, deixa de ser considerado como eficiente no modelo de múltiplos produtos, devido ao fato de aumentar em 6 vezes o número de docentes e em 5 vezes o de funcionários, o que deveria resultar numa maior capacidade de produção. Porém, isto não ocorre, registrando um baixo crescimento de cerca de 7,5% em produção. Neste contexto, este instituto foi determinado ineficiente com índice de 10%, o pior de 2000. Para os modelos não-paramétricos esta unidade apresentou comportamento semelhante, visto que o Observatório do Valongo apesar de aumentar sua quantidade de insumos, não consegue transformá-los em produtos, diminuindo assim sua eficiência média.

Figura 4 - Evolução dos insumos e dos produtos
DMU9 - Observatório do Valongo



Os modelos DEA pelas suas características e atrativos, podem ser uma alternativa de avaliação de eficiência e de possíveis melhorias no desempenho da produção. Entretanto, cabe ressaltar que a obtenção da eficiência máxima não indica que a unidade de análise está livre de qualquer melhora na sua produtividade.

Correlação entre os modelos

Neste trabalho verifica-se que as metodologias empregadas apresentam resultados semelhantes. Analisam-se tais fatos por meio de medidas de correlação dos postos associados aos níveis de eficiência dos modelos. A correlação aqui proposta advém da estatística de Spearman, a qual mede a correlação dos postos obtidos nos diferentes modelos. O quadro a seguir consta de uma matriz triangular que contém as correlações entre os postos das medidas de eficiência dos diferentes modelos abordados.

Quadro - Correlações de Spearman entre os postos associados aos níveis de eficiência dos modelos estudados

Correlações		FC			FPE			FPEH			DEA-1			DEA-M			FCT			FPEHT		
		1998	1999	2000	1998	1999	2000	1998	1999	2000	1998	1999	2000	1998	1999	2000	1998	1999	2000	1998	1999	2000
FC	1998	1.00																				
	1999		1.00																			
	2000			1.00																		
FPE	1998	0.27			1.00																	
	1999		0.35			1.00																
	2000			0.34			1.00															
FPEH	1998	0.16			0.78			1.00														
	1999		0.54			0.59			1.00													
	2000			0.58			0.59			1.00												
DEA-1	1998	0.15			0.47			0.74			1.00											
	1999		0.26			0.75			0.63			1.00										
	2000			0.20			0.71			0.62			1.00									
DEA-M	1998	0.28			0.25			0.51			0.78			1.00								
	1999		0.48			0.64			0.57			0.79			1.00							
	2000			0.51			0.56			0.57			0.50			1.00						
FCT	1998	0.87			0.39			0.25			0.24			0.37			1.00					
	1999		0.79			0.36			0.57			0.22			0.44			1.00				
	2000			0.90			0.37			0.55			0.28			0.59			1.00			
FPEHT	1998	0.41			0.68			0.77			0.74			0.58			0.53			1.00		
	1999		0.53			0.60			0.83			0.58			0.47			0.44			1.00	
	2000			0.62			0.51			0.84			0.58			0.67			0.63			1.00

Nota-se que os modelos com maiores correlações de certa maneira possuem algumas características em comum. Em geral, os modelos têm correlação acima de 50%, um bom indicativo de que os modelos apresentam similaridade entre os resultados. Porém, principalmente no ano de 1998 os modelos apresentaram correlação baixa.

Alguns trabalhos na literatura como: Coelli & Perelman (1999) e Bauer, Berger, Ferrier & Humphrey (1998) apresentam abordagens similares e revelam baixa correlação entre as metodologias em estudo (entre 0.15 e 0.30). Assim, se o intuito for classificar as unidades, a escolha do modelo passa a ser primordial neste enfoque. Esta escolha dependerá dos dados disponíveis e da sensibilidade do decisor, que deverá ser capaz de selecionar aquele que melhor traduza a realidade dos dados em termos de insumos e produtos.

4. Conclusões e extensões

A principal proposta deste artigo foi apresentar algumas das alternativas metodológicas disponíveis na literatura atual para a estimação de fronteiras de produção. Vale ressaltar que pela natureza da atividade exercida pela Universidade não cabe a utilização de funções custos, alternativa natural quando se está diante de problemas de múltiplos produtos. Foram desenvolvidos modelos paramétricos e não paramétricos para a estimação das eficiências dos Institutos da UFRJ. Dados na forma de painel, relativos aos anos de 1998, 1999 e 2000 foram utilizados, permitindo que a evolução temporal das eficiências fosse, também, avaliada.

Em geral, apesar de os três modelos estarem buscando uma fronteira de produção, os resultados finais obtidos apresentam consideráveis diferenças, ou seja, os níveis de eficiência de cada unidade de análise quando ordenados variam de modelo para modelo. Isto pode ser decorrente da importância que cada modelo atribuiu a cada variável, capacitando cada um a ter sua particularidade. Todavia, a aplicação dos três modelos possibilita identificar os institutos de Biofísica, de Macromoléculas e a Coordenação de Programas de Pós-graduação em Engenharia - COPPE como sendo os de melhor produtividade na Universidade.

Conclui-se que os modelos possibilitam avaliar a produtividade dos institutos da Universidade, ressaltando-se a flexibilidade dos modelos de fator crítico, hierárquicos, em conciliar a complexidade da estrutura da universidade e a simplicidade de compreensão dos resultados. Os modelos DEA são, também, capazes de discriminar as unidades eficientes das ineficientes. Evidentemente, a escolha da metodologia a ser considerada concerne ao tomador de decisão, pois conforme visto anteriormente, cada modelo admite uma especialidade que depende essencialmente de sua própria estrutura e natureza.

Finalmente, algumas sugestões para desenvolvimentos futuros incluem:

i. A ineficiência técnica dos modelos de fronteira de produção estocástica poderiam assumir outras distribuições, diferentes da lognormal utilizada neste trabalho, o que possibilitaria obter novas escalas de eficiência.

ii. Em relação aos resultados do modelo DEA pode-se utilizar a flexibilidade nos pesos, esta é considerada como uma das grandes vantagens do modelo DEA. Assim, deve-se estabelecer limites entre os quais os pesos podem variar, representando restrições adicionais na formulação original. A nova formulação determina a eficiência de uma unidade, a qual será menor ou igual à da formulação original. Este procedimento denomina-se modelo com restrições aos pesos em cada variável (Estellita & Meza, 2000). Isto evita que variáveis tenham peso zero, ou seja, cada variável teria seu grau de importância no processo de mensuração das medidas de eficiência.

iii. Uma alternativa interessante seria utilizar a metodologia proposta por Tsionas (2003). Esta técnica propõe uma combinação entre os modelos de fronteira de produção estocástica e os modelos de programação matemática, nos quais os resultados do modelo DEA são utilizados como *prioris* dos níveis de eficiência no modelo de fronteira de produção estocástica. Assim, a conciliação entre as informações de ambos os modelos proporcionaria um variante capaz de mensurar os níveis de eficiência.

iv. A obtenção dos ponderadores atribuídos às variáveis e da elaboração do produto agregado (*output* único) podem se constituir de um elemento subjetivo nas especificações dos modelos. Assim, seria importante um estudo adicional a fim de avaliar o impacto da atribuição subjetiva que essas especificações apresentariam sobre o resultado final.

Anexo 1

Os métodos de simulação de Monte Carlo via cadeias de Markov (MCMC), utilizado principalmente em inferência Bayesiana, têm como idéia básica a simulação de uma cadeia de Markov com distribuição limite π de forma que esta convirja para a distribuição que se tem interesse, da qual se desejam gerar amostras.

Existem vários testes para monitoração da convergência de cadeias de Markov, assim como são destacados dois algoritmos para a geração da cadeia com uma determinada distribuição estacionária π. Dentre eles, têm-se o amostrador de Gibbs e o algoritmo de Metropolis-Hastings.

As distribuições condicionais completas são dadas por:

Caso Gamma: suponha que $(u_i|\alpha,\beta) \approx Ga(\alpha/2, \beta/2)$, então as posteriores condicionais completas são:

i) $p(u_i \mid \beta, \sigma_v^{-2}, \sigma_u^{-2}, \alpha, y) \propto u_i^{\alpha/2-1} N[u_i \mid \mu_i, \sigma_u^2], onde\ \mu_i = -(y_i^* + \sigma_v^{-2})\beta/2\ e\ N[x \mid m, S]$ significa que $x \sim N[m, S]$

ii) $(\beta \mid \sigma_v^{-2}, \sigma_u^{-2}, \alpha, u_i, y) \approx N_m[b_1, B_1], onde\ b_1 = [xy^{**}/2\sigma_v^2 + B^{-1}b]B_1, e\ B_1 = [x'x/\sigma_v^2 + B_1^{-1}]^{-1}$

iii) $(\sigma_v^{-2} \mid \beta, \sigma_u^{-2}, \alpha, u_i, y) \approx Ga[c_1/2, C_1/2], onde\ c_1 = c+m, e\ C_1 = [C + (y^{**} - x\beta)'(y^{**} - x\beta)]^{-1}$

iv) $(\sigma_u^{-2} \mid \beta, \sigma_v^{-2}, \alpha, u_i, y) \approx Ga[d_1/2, D_1/2], onde\ d_1 = d + m\alpha, e\ D_1 = [D + T_2(u)]\ onde T_2(u) = \sum_{i=1}^{m} u_i$

v) $p(\alpha \mid \beta, \sigma_v^{-2}, \sigma_u^{-2}, u_i, y) \propto (\beta^{\alpha/2}/\Gamma(\alpha/2))^m T_1(u)^{\alpha/2-1} Ga(d/2, D/2), onde T_1(u) = \prod_{i=1}^{m} u_i$

Caso Lognormal: suponha $(u_i|\alpha,\sigma^2) \approx LN(\mu, \sigma_u^2)$, então as posteriores condicionais completas são:

i) $p(u_i \mid \beta, \sigma_v^{-2}, \sigma_u^{-2}, \mu, y) \propto N[y_i^*, \sigma_v^2] * LN[\mu, \sigma_u^2]$

ii) $(\beta \mid \sigma_v^{-2}, \sigma_u^{-2}, \mu, u_i, y) \approx N_m[b_1, B_1], onde\ b_1 = [xy^{**}/\sigma_v^2 + B^{-1}b]B_1, e\ B_1 = [x'x + B_1^{-1}]^{-1}$

iii) $(\sigma_v^{-2} \mid \beta, \sigma_u^{-2}, \mu, u_i, y) \approx Ga[c_1/2, C_1/2], onde\ c_1 = c+m, e\ C_1 = [C + (y^{**} - x\beta)'(y^{**} - x\beta)]^{-1}$

iv) $(\sigma_u^{-2} \mid \beta, \sigma_v^{-2}, \mu, u_i, y) \approx Ga[d_1/2, D_1/2], onde\ d_1 = d+m, e\ D_1 = [D + \sum_{i=1}^{m} [\log(u_i - \mu)^2]^{-1}$

v) $(\mu \mid \beta, \sigma_v^{-2}, \sigma_u^{-2}, u_i, y) \approx N(a_1, A_1),\ onde\ a_1 = \left[u^* m\sigma_u^2 + A_1^{-1}a\right]A_1$, e $A_1 = [m\sigma_u^{-2} + A^{-1}]^{-1}$

As distribuições condicionais completas anteriores não estão disponíveis em uma forma fechada e assim, o amostrador de Gibbs não pode ser implementado. Porém, como as distribuições condicionais dos teoremas 2 (itens i e iv) e 3 (item i) são logconcavas, é possível utilizar métodos de amostragem de rejeição adaptativa, no qual utiliza-se o algoritmo de Metropolis-Hastings (Chib & Geenberg, 1995).

Anexo 2

O critério de minimização do desvio preditivo esperado foi introduzido por Gelfand & Ghosh (1998), no qual uma decisão d deve ser selecionada para simultaneamente estimar as quantidades observadas y_0 e predizer algumas replicações dos dados anteriores y_r. Para cada componente de y_r tem-se uma decisão d_i, e segue uma função de perda composta por:

$$L(y_{ir}, d_i; y_{i0}) = L(y_{ir}, d_i) + kL(y_{i0}, d_i) \text{ onde } k \geq 0.$$

Quando o peso k é fixo e igual a zero e o y_r é de fato uma observação nova, a decisão d corresponde a escolha de uma predição para uma observação futura. Alternativamente, a funcão de perda proposta de decisão escolhida ótima poderia acomodar também a proximidade de y_{i0}. Usando uma função de perda quadrática e agregando as componentes do vetor y_r, o minimo esperado do desvio quadrático é determinado por:

$$D_k(M) = \sum_{i=1}^{n} \sigma_i^{2(M)} + \frac{K}{K+1} \sum_{i=1}^{n} (\mu_i^{(M)} - y_{i0})^2.$$

Esta expressão pode ser rescrita por:

$$D_k(M) = P(m) + \frac{K}{K+1} G(m)$$

onde o termo *G(m)* é a soma dos erros ao quadrado, uma medida da adequação do ajuste, e *P(m)* pode ser interpretado como uma penalidade à maior complexidade do modelo. A medida que os modelos ficam mais complexos, o termo *G(m)* diminui, enquanto *P(m)* aumenta. Geralmente, a escolha do modelo mais parcimonioso acaba sendo estimulada, penalizando a complexidade do modelo.

Um caso limite, onde k $\to \infty$, denota-se por minimização do desvio preditivo esperado, EPD, neste caso: *D(M) = P(m) + G(m)*: Quanto menor o valor de EPD, melhor o modelo.

Anexo 3

Pesos Escolhidos para a construção dos indicadores

DOCENTES			
	GRADUADOS	**MESTRADO**	**DOUTORADO**
20 HORAS	0,4	0,5	1,25
40 HORAS	0,8	1,0	2,5
DE	1,2	0,5	3,75

Índice de Publicação		
Artigos periódicos nacionais		**24**
Artigos periódicos internacionais		**30**
Congressos Nacionais		
	Completo	8
	Resumo	2
Congressos Internacionais		
	Completo	16
	Resumo	14
Outras publicações: apresentações, seminários, conferências, exposições, etc.		1

Índice de Orientação	
Mestrado	4
Doutorado	6

Referências bibliográfica

AIGNER, D., LOVELL, C.A. K, SCHMIDT (1977), *"Formulation and estimation of stochastic frontier production function models"*, Journal of Econometrics, v.6, p.21-37.

BANKER, R.D., DATAR, S.M. E KEMERER, C.F. (1991), *"A model to evaluate variables impacting the productivity of software maintenance projects"*, Management Science, v. 37, p. 1-18.

BANKER, R.D., CHARNES A., COOPER W.W. (1984), *"Soma models for estimating technical and scale inefficiencies in Data Envelopment Analysis"*. Management Science, v.30, n.9, p.1078-1092.

BAUER, P.W. (1990), *"Recent developments in the econometric estimation of frontier"*, Journal of Econometrics, v.47, p.39-56.

BAUER, P.W., BERGERN, A.N., FERRIER, G.D., HUMPHREY, D.B. (1998), *"Consistency conditions for regulatory analysis of financial institutions: a comparison of frontier efficiency methods"*. North Holland.

BROOKS, S.P., GELMAN, A. (1998), *"General methods for monitoring convergence of iterative simulations"*. Journal of Computational and Graphical statistic, v.7, n. 4, p. 434-455.

BRYK, A.S., RAUDENBUSH, S.W. (2002), *"Hierarchical Linear Models"*, Sage Publications, Newbury Park, CA, second edition.

CASELLA, G., GEORGE, E.I. (1992), *"Explaining the gibbs sampler"*, American Statistical Association, v.46, n. 3, p. 167-174.

CHARNES, A.,W.W. COOPER E RHODES (1978), *"Measuring the efficiency of decision making units"*, European Journal of Operational Research , v.2,n. 6, p.429-444.

CHIB, S., GREENBERG, E. (1995), *"Understanding the Metropolis-Hastings Algorithm"*, The American Statistician, v. 49, n.4, pp. 327-335.

COELLI, T., PERELMAN, S. (1999), *"A comparison of parametric and nonparametric distance functions: With application to European railways"*. European Journal of Operational Research.

ESTELLITA, M. P., MEZA, L. A. (2000), *"Análise envoltória de dados e perspectivas de integração no ambiente do apoio à decisão"*. Rio de Janeiro: COOPE/UFRJ.

FORSUND, F.R., LOVELL, C.A.K., SCHMIDT,P. (1980), *"A survey of frontier production frontiers and their relationship to efficiency measurement"*, Journal of Econometrics, v.13, p.5-25.

Gamerman, D., Migon, H.S., Sant'ana, A.P. (1992), *"Um modelo integrado para melhoramentos da qualidade das universidades públicas"*, Relatório técnico n.65, Instituto de Matemática, Universidade Federal do Rio de Janeiro.

GAMERMAN, D. (1997) *"Markov Chain Monte Carlo"*, ChapmanHall.

GELFAND, A. E., GHOSH, S. K. (1998), *"Model Choice: A minimum posterior predictive loss approach"*, Biometrika, v.85, 1-11.

GILKS, W. R., RICHARDSON, S. AND SPIEGELHALTER, D. J. (1995), *"Markov Chain Monte Carlo in Practice"*, Chapman-Hall.

GREENE,W.I.L. (1993), *"The econometric approach to efficiency analysis"* New York: Oxford University Press, p.68-119.

JONHSON, R.A.,WICHERN, D.W. (1999), *"Applied Multivariate Statistical Analysis"*, Prentice Hall, fourth edition.

MEEUSEN, W., BROECK, J. VAN DEN (1977), *"Efficiency estimation from Cobb-Douglas production with composed error"*, International Economics Review, v. 18, n. 2, p. 435-444.

MEDRANO L.A, MIGON, H.S. (2003), *"Seleção de modelos de fronteira de produção estocástica: abordagem bayesiana"*, Rel.Texc DME-IM/UFRJ.

MIGON, H.S., GAMERMAN, D. (1999). *"Statistical Inference: An Integrated Approach"*. London. Arnold

MIGON, M.N. (2000), *"Eficiência da indústria do transporte aéreo no Brasil: uma avaliação de Análise Envoltória de Dados (DEA)"*. Rio de Janeiro: COPPE/UFRJ, dissertação de mestrado.

MIGON, H. S. (2001) *"A Bayesian Approach to Stochastic Production Frontier"*, Tech. Report n. 105 , DME-IM/UFRJ.

SCHIMIDT, P. (1986), *"Frontier production functions"*, Econometric Reviews, v.4, p.289-328.

SPIEGELHALTER, D. J. et all (1996), *"BUGS Bayesian Analysis using Gibbs Sampling"*, version 0.5.

TISIONAS, E.G. (2003), *"Combining DEA and stochastic frontier models: An emprirical Bayes approach"* , European Journal of Operational Research, n. 147, p.499-510

Universidade Federal do Rio de Janeiro, UFRJ (2000), *"Relatório de gestão"*.

Agradecimentos

Os autores agradecem aos revisores e ao editor responsável pelas sugestões que contribuíram para a melhoria deste trabalho. Além disto, destacam o valioso apoio das agências de financiamento: Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Capes.

Abstract

In this study a revision of models of stochastic frontiers production is presented. Among the approached problems we point out: the choice of the distribution to represent the efficiencies, the modelling of panel data and the available methodological alternatives for the treatment of situations with multiple products. The introduced models are members of the class of the Bayesian hierarchical models and the inferences will be performed throught methods of stochastic simulation - Markov Chain Monte Carlo methods. Thus, the relative effects of the inputs about the production -elasticities- may be modeled as specific firm or not. The models revised in this article are applied to real data relative to the Federal University of Rio de Janeiro (UFRJ) and evaluated in three periods of time. The obtained results are annalysed and ilustrated, highlightning the potentialities of those techniques as useful tools to take decisions.

REVISTA BRASILEIRA DE ESTATÍSTICA - RBEs

POLÍTICA EDITORIAL

A Revista Brasileira de Estatística - RBEs publica trabalhos relevantes em Estatística aplicada, não havendo limitação no assunto ou matéria em questão. Como exemplos de áreas de aplicação citamos as áreas de advocacia, ciências físicas e biomédicas, criminologia, demografia, economia, educação, estatísticas governamentais, finanças, indústria, medicina, meio ambiente, negócios, políticas públicas, psicologia e sociologia, entre outras. A RBEs publicará também artigos abordando os diversos aspectos de metodologias relevantes para usuários e produtores de estatísticas públicas, incluindo planejamento, avaliação e mensuração de erros em censos e pesquisas, novos desenvolvimentos em metodologia de pesquisa, amostragem e estimação, imputação de dados, disseminação e confiabilidade de dados, uso e combinação de fontes alternativas de informação e integração de dados, métodos e modelos demográficos e econométricos.

Os artigos submetidos devem ser inéditos e não devem ter sido submetidos simultaneamente a qualquer outro periódico.

O periódico tem como objetivo a apresentação de artigos que permitam fácil assimilação por membros da comunidade em geral. Os artigos devem incluir aplicações práticas como assunto central, com análises estatísticas exaustivas e apresentadas de forma didática. Entretanto, o emprego de métodos inovadores, apesar de ser incentivado, não é essencial para a publicação.

Artigos contendo exposição metodológica são também incentivados, desde que sejam relevantes para a área de aplicação pela qual os mesmos foram motivados, auxiliem na compreensão do problema e contenham interpretação clara das expressões algébricas apresentadas.

A RBEs tem periodicidade semestral e também publica artigos convidados e resenhas de livros, bem como incentiva a submissão de artigos voltados para a educação estatística.

Artigos em espanhol ou inglês só serão publicados caso nenhum dos autores sejam brasileiros e nem residam no País.

Todos os artigos submetidos são avaliados quanto à qualidade e à relevância por dois especialistas indicados pelo Comitê Editorial da RBEs.

O processo de avaliação dos artigos submetidos é do tipo 'duplo cego', isto é, os artigos são avaliados sem identificação de autoria e os comentários dos avaliadores também são repassados aos autores sem identificação.

INSTRUÇÃO PARA SUBMISSÃO DE ARTIGOS À RBEs

O processo editorial da RBEs é eletrônico. Os artigos devem ser submetidos via email para: hortega@ibge.gov.br

Após a submissão o autor correspondente receberá um código para acompanhar o processo de avaliação do artigo. Caso não receba um aviso com este número no prazo de uma semana, fazer contato com a secretaria da revista no endereço:

Revista Brasileira de Estatística
IBGE - Diretoria de Pesquisas - Coordenação de Métodos e Qualidade
Av. República do Chile, nº 500, 10º andar
Centro, Rio de Janeiro - RJ
CEP: 20031-170
Tel.: 55 21 2142-0472
55 21 2142-4549
Fax.: 55 21 2142-4802

INSTRUÇÃO PARA PREPARO DOS ORIGINAIS

Os originais entregues para publicação devem obedecer as normas seguintes.

1. Originais processados pelo editor de textos Word for Windows são preferidos. Entretanto, serão aceitos também originais processados em LaTeX desde que sejam encaminhados acompanhados de versões em pdf, conforme descrito no item 3 a seguir;

2. A primeira página do original (folha de rosto) deve conter o título do artigo, seguido do(s) nome(s) completo(s) do(s) autor(es), indicando-se, para cada um, a afiliação e endereço para correspondência. Agradecimentos a colaboradores e instituições, e auxílios recebidos também devem figurar nesta página;

3. No caso da submissão não ser em Word for Windows, três arquivos do original devem ser enviados. O primeiro deve conter os originais no processador de texto utilizado (por exemplo, Latex). O segundo e terceiro devem ser no formato pdf, sendo um com a primeira página como descrito no item 2 e outro contendo apenas o título, sem identificação do(s) autor(es) ou outros elementos que possam permitir a identificação da autoria;

4. A segunda página do original deve conter resumos em português e inglês (abstract), destacando os pontos relevantes do artigo. Cada resumo deve ser digitado seguindo o mesmo padrão do restante do texto, em um único parágrafo, sem fórmulas, com, no máximo, 150 palavras;

5. O artigo deve ser dividido em seções, numeradas progressivamente, com títulos concisos e apropriados. Todas as seções e subseções devem ser numeradas e receber título apropriado;

6. Tratamentos algébricos exaustivos devem ser evitados ou alocados em apêndices;

7. A citação de referências no texto e a listagem final de referências devem ser feitas de acordo com as normas da ABNT;

8. As tabelas e gráficos devem ser precedidos de títulos que permitam perfeita identificação do conteúdo. Devem ser numeradas seqüencialmente (Tabela 1, Figura 3, etc.) e referidas nos locais de inserção pelos respectivos números. Quando houver tabelas e demonstrações extensas ou outros elementos de suporte, podem ser empregados apêndices. Os apêndices devem ter título e numeração, tais como as demais seções de trabalho; e

9. Gráficos e diagramas para publicação devem ser incluídos nos arquivos com os originais do artigo. Caso tenham que ser enviados em separado, devem ter nomes que facilitem a sua identificação e posicionamento correto no artigo (ex.: Gráfico 1; Figura 3; etc.). É fundamental que não existam erros, quer no desenho, quer nas legendas ou títulos.

Se o assunto é **Brasil**,
procure o **IBGE**

www.ibge.gov.br
wap.ibge.gov.br

atendimento
0800 218181