revista brasileira de estudos de população

avanços metodológicos recentes na medição dos diferenciais da mortalidade

Ralph Hakkert*

RESUMO - Discutem-se algumas técnicas recentes que visam facilitar a análise da mortalidade diferencial segundo diversos critérios simultâneos, em situações onde as amostras pequenas impedem um cruzamento por todas as dimensões relevantes. Todas estas técnicas envolvem o uso de modelos de regressão. Embora alguns detalhes técnicos sejam inevitáveis, visa-se principalmente providenciar uma introdução ao assunto para os não-especialistas, com uma ampla bibliografia dos trabalhos teóricos e aplicados atualmente desenvolvidos na área. O raciocínio subjacente ao método dos riscos proporcionais de Cox, o mais comum nas aplicações atuais, é explicado, e suas vantagens e desvantagens são apontadas. Levantam-se alternativas mais apropriadas para a análise da mortalidade em populações humanas. Discutem-se ainda o problema da heterogeneidade não explicada e algumas extensões para a estimação indireta, inspiradas pelo método de Cox.

INTRODUÇÃO

Na sua forma mais típica, a demografia formal estuda fenômenos de escala macrossocial. Procura-se observar, descrever e sistematizar dados referentes à estrutura e dinâmica de grandes conjuntos de indivíduos através de generalizações estatísticas expressadas em termos de somas, proporções ou taxas. O percentual de mulheres em idade fértil, a razão de sexo, a soma dos indivíduos em idade ativa, o risco de morte, a taxa de natalidade são conceitos gerais referentes a grandes universos que, atrás da sua aparente objetividade,

* Núcleo de Estudos de População (NEPO) da Universidade Estadual de Campinas e Grupo Especial de Análise Demográfica (GEADE) da Fundação SEADE, São Paulo.

escondem uma considerável variação de comportamentos individuais e relações sociais. Em certos casos, podem ser entendidos simplesmente como tendências centrais de processos que, sabidamente, possuem intensidades mais ou menos heterogêneas. Noutros, porém, a descrição estatística implicitamente pressupõe que os grupos sociais descritos sejam homogêneos em relação às características demográficas em questão. Sendo assim, o não-reconhecimento da heterogeneidade existente pode causar vieses apreciáveis e afetar inclusive a interpretação de determinadas estatísticas vitais. Noutros casos ainda, a variação sistemática dos fenômenos demográficos em relação a características sociais, econômicas, ou biológicas como a educação, a renda, ou o peso ao nascer têm um interesse substantivo próprio.

O objetivo do atual artigo é comentar alguns desenvolvimentos metodológicos recentes que visam facilitar o estudo das possíveis conseqüências e, principalmente, das origens desta heterogeneidade. Todos usam modelos de regressão, especialmente de regressão binária ou categórica. Embora algumas explicações técnicas sejam inevitáveis, visa-se principalmente uma introdução sistemática ao assunto para o não-especialista. Uma ampla bibliografia, com trabalhos recentes na área, tanto aprofundamentos técnicos quanto aplicações práticas, é anexada. As fórmulas que aparecem em algumas das secções servem para a maior precisão dos conceitos teóricos discutidos no texto, mas não são essenciais numa primeira leitura.

1. DOIS TIPOS DE ABORDAGEM

Enquanto os principais avanços metodológicos dos anos 60 e 70 consistiram no melhoramento qualitativo dos indicadores demográficos globais, através das chamadas técnicas indiretas, um dos problemas que ultimamente têm recebido mais atenção é a análise da diversidade inerente que as taxas globais não conseguem transmitir. Um dos fatores que vêm contribuindo para esta mudança de ênfase é a disponibilidade crescente de levantamentos específicos, como a Pesquisa Mundial de Fecundidade, que freqüentemente contêm histórias de vida completas, dispensando, portanto, a necessidade de estimativas indiretas. Por outro lado, o objetivo de estudar as interrelações e os determinantes sócio-econômicos destas taxas com amostras inerentemente pequenas exige a adoção de técnicas estatísticas específicas.

De modo geral, os esforços que vêm sendo despendidos no assunto podem ser divididos em duas categorias.

revista
brasileira de
estudos de
população

1.1 O ESTUDO DOS EFEITOS DA HETEROGENEIDADE NÃO EXPLICADA

Qualquer aspecto da dinâmica demográfica afeta os indivíduos com intensidades distintas. A não ser que estas variações e seus determinantes sociais, econômicos ou biológicos constituam o próprio objeto da análise, são geralmente consideradas neutras, no sentido de que as taxas globais resumem o aspecto mais importante do processo analisado, ou seja, a sua tendência central. Por exemplo, a constatação de uma queda na taxa de natalidade pode ser suficiente como conclusão de determinado tipo de inquérito demográfico, não importando - neste nível de análise - se esta queda é uniforme ou causada por mudanças no comportamento reprodutivo de apenas parte da população. Acontece, entretanto, que certas propriedades formais dos indicadores demográficos dependem essencialmente da homogeneidade ou heterogeneidade dos fenômenos estudados. Por exemplo:

- O padrão etário de uma tábua de vida depende, entre outros fatores, da variação dos riscos de morte entre as diversas camadas da população. Analiticamente, pode ser demonstrado que uma heterogeneidade maior tende a elevar a mortalidade proporcional nas faixas mais jovens (Manton, Stallard & Vaupel, 1981; Vaupel, Manton & Stallard, 1979; Vaupel & Yashin, 1983);
- A variação aleatória a ser esperada na taxa anual de natalidade de uma área relativamente pequena é maior na medida em que o risco de engravidar for mais igualmente distribuído entre todas as mulheres em idade fértil (Avery & Hakkert, 1981);
- Um país com grandes desigualdades na distribuição dos recursos sociais e econômicos determinantes do nível da mortalidade tenderá, mesmo por razões puramente analíticas, a ter uma esperança de vida mais baixa do que outro, com recursos equivalentes, mas divididos de forma menos desequilibrada (Rodgers, 1979);
- Onde coexistem níveis de mortalidade muito díspares, o efeito de uma redução das taxas de mortalidade sobre a esperança de vida depende de como esta é distribuída entre as diversas camadas da população (Keyfitz & Littman, 1980; Shepard & Zeckhauser, 1975, 1977);
- Finalmente, as técnicas indiretas usuais na medição da mortalidade infantil e de crianças supõem implicitamente que a distribuição da fecundidade por idade seja igual para todas as mulheres. Caso contrário, as estimativas resultantes contêm um viés que se agrava na medida em que houver alguma correlação entre os níveis da fecundidade e mortalidade dos diversos setores da população.

A característica que diferencia as contribuições nesta categoria das que serão discutidas em seguida é a preocupação com os efeitos analíticos da heterogeneidade, independentemente das suas origens. Embora algumas das questões levantadas por esta literatura tenham implicações importantes, o atual trabalho dará mais ênfase a segunda categoria, que enfoca justamente a origem das variações observadas.

1.2 O ESTUDO DE ASSOCIAÇÕES ENTRE AS TAXAS DEMOGRÁFICAS E OUTROS FATORES

A linha de pesquisa a ser discutida aqui, e que forma o assunto principal do atual artigo, tem uma história longa nos países industrializados, embora as limitações na disponibilidade de dados façam com que sua aplicação nos países em desenvolvimento, na maioria dos casos, seja mais recente. Trata-se da decomposição de taxas demográficas por setores da população (definidos em termos de características relevantes do ponto de vista do processo estudado, como renda, educação, ocupação ou outras), seja com o objetivo de encontrar pistas de possíveis relações causais, seja com a intenção mais modesta de identificar os grupos populacionais mais intensamente afetados. Podem ainda ser distinguidas a chamada abordagem ecológica, que se baseia na comparação de unidades geográficas em termos destas características (por exemplo, Santos et al., 1976; Singer et al., 1978), e análises onde os agrupamentos são realizados diretamente em termos das variáveis estudadas (por exemplo, Hakkert, 1984; Simões, 1981). O atual trabalho se limitará ao segundo procedimento. Embora alguns dos métodos discutidos se adaptem também ao estudo de outros fenômenos demográficos, a ênfase será dada a sua aplicação para a medição dos diferenciais da mortalidade.

2. DEFINIÇÃO DO PROBLEMA

Deixando de lado um dos obstáculos metodológicos tradicionais ao estudo da mortalidade diferencial nos países industrializados, ou seja, como chegar a uma classificação uniforme dos eventos analisados e da população a eles exposta, parece mais apropriado enfocar diretamente as complicações da abordagem por meio de "surveys". Ao contrario do que acontece nos estudos tradicionais que obtêm dados sobre eventos vitais através do registro civil, relacionando estes com uma classificação da população correspondente, calculada na base de um censo (Fox, 1979; Kitagawa & Hauser, 1973), a metodologia mais comum nos estudos que atualmente estão sendo realizados nos países em desenvolvimento evita esta complicação. Usando uma única fonte retrospectiva - seja censo, seja um

levantamento específico -, são classificados tanto o número de eventos quanto a população exposta. No caso de dados censitários, o principal problema diz respeito à escassez das informações sobre eventos vitais que podem ser extraídas do material. Esta leva inevitavelmente à aplicação de métodos indiretos que - além de sempre sujeitos a dúvidas quanto a possíveis vieses - são estatisticamente pouco eficientes. Entretanto, na medida em que - até há poucos anos - dados censitários constituíam quase a única fonte facilmente disponível, a maioria dos estudos até agora realizados, como os trabalhos de Hugo Behm e sua equipe (Behm et al., 1976-1980), tem se baseado nesta metodologia. No final do atual artigo serão feitas algumas referências a desenvolvimentos recentes nesta área. A questão principal, porém, é como aproveitar os dados coletados em levantamentos específicos, como a Pesquisa Nacional sobre a Reprodução Humana, do CEBRAP, ou a Pesquisa Mundial de Fecundidade. Pesquisas como estas, baseadas em "surveys", geralmente apresentam a vantagem de conter informações mais completas sobre os eventos vitais em questão. Entretanto, na medida em que seus tamanhos amostrais raramente superam a faixa dos 10.000 domicílios, a variância aleatória das estimativas obtidas torna-se uma preocupação dominante.

Para esclarecer o assunto, imagine uma análise hipotética visando aferir os diferenciais da mortalidade infantil entre 9 setores da população, por exemplo, 3 faixas de renda cruzadas com 3 níveis educacionais. Supondo que a TMI média seja da ordem de 75 por 1.000 e a diferença mínima entre as taxas setoriais da ordem de 10 por 1.000, precisa-se - no mínimo - de umas 45.000 histórias de vida de crianças para poder concluir, com margem de erro inferior a 5%, que todos os diferenciais observados sejam estatisticamente significativos. Com esquemas de classificação mais complexos, as exigências dos tamanhos amostrais mínimos crescem rapidamente, inviabilizando cruzamentos de mais de duas ou - no máximo - três variáveis simultâneas. Evidentemente, esta é uma situação pouco satisfatória, considerando o desejo de aproveitar estes levantamentos para avaliar isoladamente o impacto de cada variável pertinente, na ausência de perturbações por parte de outros fatores. Surge, portanto, a necessidade de um método estatístico que consiga resumir o impacto de cada variável sem levar a uma diluição da amostra em intermináveis sub-divisões.

Na falta de condições para analisar cada sub-divisão da amostra separadamente, a alternativa lógica é impor algum tipo de modelo matemático à estrutura dos dados, trocando as-

sim a flexibilidade dos cruzamentos pela maior eficiência estatística do modelo. Evidentemente, esta troca não deixa de ter suas desvantagens. Na medida em que os dados apresentam surpresas não previstas na estrutura do modelo, existe o perigo de que certos aspectos interessantes da dinâmica subjacente passem desapercebidos. Também, a própria complexidade computacional de alguns dos procedimentos pode formar um obstáculo à sua aplicação. Entretanto, é importante enfatisar que a alternativa, ou seja, analisar apenas o impacto marginal de uma ou duas variáveis de cada vez, apresenta um grande perigo quando o objetivo é não apenas descrever a desigualdade existente, mas também tirar alguma conclusão sobre a estrutura causal que determina os diferenciais observados. Principalmente quando as variáveis descritivas são altamente correlacionadas, como no caso da educação e da renda, corre-se o risco de atribuir um valor explicativo exagerado à variável que apresenta o mais alto poder discriminatório.

Um procedimento idôneo para analisar o impacto de diversas variáveis independentes (no caso, educação, renda, peso ao nascer, etc.) sobre a variável dependente (no caso, algum indicador do nível da mortalidade) é a regressão, onde se expressa o impacto das variáveis $y_1,\ldots,y_p$ (variáveis independentes, ou transformações e combinações destas) sobre a variável dependente D em termos de alguma transformação f e uma combinação linear y de $y_1,\ldots,y_p$, ou seja:

$$D = f(y) = f(\beta_o + \beta_1 y_1 + \ldots\ldots + \beta_p y_p) + E \qquad (1)$$

onde E representa um erro aleatório, e o objetivo da análise é avaliar os coeficientes $\beta_o,\ldots,\beta_p$. No caso mais direto e mais conhecido, $y_1,\ldots,y_p$ são simplesmente as p variáveis substantivas cujo impacto está sendo analisado, e f é a transformação unitária, de modo que se obtém a equação da regressão linear comum, ou seja:

$$D = y = \beta_o + \beta_1 y_1 + \ldots\ldots + \beta_p y_p + E \qquad (2)$$

Noutros casos, porém, é necessário realizar alguma transformação, quadrática, logarítmica, ou seja qual for, para converter y em D. Tanto esta transformação f quanto a relação entre $y_1,\ldots,y_p$ e as variáveis substantivas devem ser especificadas de antemão. Por exemplo, se suspeita-se que o impacto do peso ao nascer sobre a mortalidade infantil é quadrático, é preciso incluir dois termos entre $y_1,\ldots,y_p$ para expressar esta relação: o peso ao nascer e o seu quadrado. Se não, é possível que a análise não registre qualquer efeito aparente desta variável.

Fora estas advertências gerais, que não constituem nenhuma novidade, a tarefa de aplicar um procedimento de regressão à análise da mortalidade diferencial enfrenta um problema específico, ou seja, como definir a variável dependente D? Todos os indicadores demográficos conhecidos, como a esperança de vida, a TMI ou a taxa de mortalidade por idade, são generalizações estatísticas baseadas, necessariamente, em universos inteiros, sem significado para indivíduos isolados, a não ser no sentido probabilístico. É por causa desta dificuldade fundamental que, até há poucos anos, o uso de métodos de regressão nesta área se limitava a estudos ecológicos onde tanto as variáveis independentes quanto a dependente se referem a unidades geográficas e não a indivíduos. Apesar dos perigos desta metodologia (ver Robinson, 1950; Goodman, 1953), parecia não haver uma alternativa viável.

3. O MÉTODO DOS RISCOS PROPORCIONAIS

Já faz mais de dez anos que um grupo de bio-estatísticos, confrontados com um problema muito parecido, ou seja, como separar o impacto de diversos fatores clínicos sobre as chances de sobrevivência de pacientes com doenças graves, desenvolveu o chamado modelo dos riscos proporcionais. Formulado por Cox (1972), o método recebeu algumas contribuições por parte de outros autores (por exemplo, Breslow, 1975), e atualmente já faz parte dos livros de texto na área de bio-estatística e análise de atrito de materiais (por exemplo, Elandt-Johnson & Johnson, 1980; Kalbfleisch & Prentice, 1980; Lawless, 1982). Uma introdução mais compacta e mais voltada para aplicações sociológicas pode ser encontrada em Tuma (1982). A estréia desta metodologia na demografia é mais recente, mas desde 1980 apareceram vários artigos que aplicam o método, não apenas como recurso para a análise dos diferenciais da mortalidade (Chackiel, 1982; Hobcraft et al., 1982; Martin et al., 1983; Trussell & Hammerslough, 1983), mas também para outros fenômenos, como nupcialidade e reprodução (Michael & Tuma, 1983; Trussell & Bloom, 1983), migração (Sandefur & Scott, 1981), dissolução de casamentos (Menken et al., 1981) ou relações entre a participação na força de trabalho e a fecundidade (Wynam, 1981).

Como a maioria dos modelos que serão discutidos em seguida, o método dos riscos proporcionais se baseia numa técnica muito conhecida da estatística matemática, que é o método da máxima verossimilhança. Este consiste em definir a probabilidade dos acontecimentos observados em termos dos parâmetros $\beta_0,\ldots,\beta_p$, para, em seguida, escolher os valores destes parâmetros de tal forma que esta probabilidade seja maxi-

mizada (por exemplo, ver Meyer, 1969, Cap. 14). No caso do modelo dos riscos proporcionais, o princípio é aplicado da seguinte forma.

Supõe-se que todo indivíduo pode ser caracterizado por uma função de risco $\mu(x)$, também chamada a "força da mortalidade", que expressa a probabilidade instantânea de morrer com idade x, tendo sobrevivido até então. Esta função - é claro - varia entre os indivíduos, mas de forma proporcional, ou seja:

$$\mu(x) = \lambda \mu_o(x) \qquad (3)$$

onde $\mu_o(x)$ é igual para todos, e apenas λ varia. Uma maneira equivalente, que não usa o conceito de riscos instantâneos, para expressar a mesma relação é a seguinte:

$$\log \ell_x = \lambda \log \ell_x^o \qquad (4)$$

onde ℓ_x tem seu significado usual de probabilidade de sobrevivência desde o nascimento até a idade x. Como em (3), ℓ_x^o é uma função padrão supostamente igual para todos, enquanto λ expressa a variabilidade individual dos níveis da mortalidade.

O próximo suposto é que λ pode ser escrito como uma função das variáveis independentes, como em (1). Normalmente, a transformação f adotada para este fim é a função exponencial:

$$\lambda = f(y) = \exp(\beta_1 y_1 + \ldots\ldots + \beta_p y_p) \qquad (5)$$

Nota-se a ausência do termo E expressivo do erro aleatório, uma das diferenças entre a regressão comum e o modelo dos riscos proporcionais. Nota-se também a ausência do termo β_o na parte linear da expressão que torna-se desnecessária, na medida em que pode ser considerado incorporado em $\mu_o(x)$ ou ℓ_x^o.

Uma vez especificado o modelo, segue-se o seguinte raciocínio. Suponha que o primeiro óbito ocorra na idade x_1. Entre todos os indivíduos que estavam expostos ao risco de morrer naquele momento, qual teria sido a probabilidade de que quem morreu foi justamente o indivíduo cujo óbito foi efetivamente constatado, e que ficará identificado aqui através do índice k_1? Pois bem, não é difícil ver que esta probabilidade condicional pode ser escrita como:

$$\mu_{k_1}(x_1) \Big/ \sum_{k\varepsilon\mathcal{S}(x_1)} \mu_k(x_1) = \exp(y_{k_1}) \Big/ \sum_{k\varepsilon\mathcal{S}(x_1)} \exp(y_k) \quad (6)$$

onde o conjunto $\mathcal{S}(x_1)$ se refere a todos os indivíduos presentes na idade x_1. Nota-se - e esta é uma propriedade fundamental do modelo - que a expressão (6) independe totalmente de $\mu_0(x_1)$. Por causa desta propriedade, o modelo dos riscos proporcionais é considerado um procedimento "semi-paramétrico", ou seja, depende da especificação de λ em termos de $y_1,\ldots,y_p$, mas não é preciso especificar qualquer função de risco subjacente.

O mesmo cálculo pode ser repetido no instante x_2 do segundo óbito, com a diferença de que agora o conjunto $\mathcal{S}(x_2)$ exclui o indivíduo k_1, que já morreu antes, e eventualmente pode excluir quem saiu da amostra por outras razões, como pode também incluir novas entradas que chegaram com idades entre x_1 e x_2. Continua-se desta forma até o último óbito observado. No final, multiplicam-se as diversas probabilidades condicionais para obter a função de verossimilhança L ou - computacionalmente mais conveniente - o logaritmo desta:

$$\log L = \sum_{i\varepsilon\mathcal{O}} \left(y_{k_i} - \log \sum_{k\varepsilon\mathcal{S}(x_i)} \exp(y_k)\right) \quad (7)$$

onde o conjunto $\mathcal{O}$ se refere a todos os óbitos observados. A estimação dos parâmetros $\beta_1,\ldots,\beta_p$ agora se resume em encontrar os valores que maximizem esta expressão. Computacionalmente, esta é uma tarefa algo complicada, para a qual existem, atualmente, vários programas de computador. O pacote SASS contém um procedimento para executar análises deste tipo. Também existe o programa RATES, da autoria de Nancy Tuma, da Universidade de Stanford (Tuma et al., 1979).

No caso especial onde todas as variáveis são categóricas, é possível demonstrar (Laird & Olivier, 1981) que a análise de riscos proporcionais é logicamente equivalente a uma análise de tabelas de contingência através do modelo log-linear (ver Bishop et al., 1975). Esta metodologia, que já faz parte do arsenal de técnicas comumente usado nas ciências sociais, conta com alguns programas, dos quais os melhores são provavelmente LOGLIN, desenvolvido pela Faculdade de Ciências Médicas da Universidade de Harvard (Olivier & Neff, 1976), e GLIM, da Associação Real de Estatística da Inglaterra (Baker & Nelder, 1978). Dois estudos de diferenciais de mortalidade que seguem esta linha são os de Frenzen e Hogan (1982) e de Gortmaker (1979).

4. ALGUMAS CRÍTICAS AO MODELO DE RISCOS PROPORCIONAIS

Nas aplicações para as quais o modelo dos riscos proporcionais foi originalmente desenvolvido, uma das suas principais vantagens é a propriedade semi-paramétrica que dispensa o pesquisador da obrigação de escolher uma função de risco subjacente. Sem dúvida esta propriedade é um prêmio quando analisa-se a mortalidade de uma sub-população tão específica quanto a de pacientes com condições clínicas graves, onde a especificação de tal função não é nada óbvia. Entretanto, ela tem um preço. Nota-se que todo o raciocínio desenvolvido na secção anterior gira em volta de probabilidades condicionais; as idades $x_1,\ldots,x_n$ onde os óbitos ocorrem não entram na expressão (7). É por causa desta característica que (7) é normalmente chamada a função de verossimilhança parcial, isto é, condicionada nos valores $x_1,\ldots,x_n$. Houve, inclusive, alguma dúvida inicial entre os especialistas se as estimativas assim obtidas teriam as mesmas propriedades desejáveis que normalmente as estimativas de máxima verossimilhança possuem, até que Cox (1975) conseguiu provar estas propriedades por outros caminhos. A característica semi-paramétrica está fundamentalmente ligada ao uso de probabilidades condicionais. Praticamente, esta opção se traduz numa maior variância das estimativas do que seria o caso num modelo paramétrico, com μ_0 explicitado desde o início. Noutras palavras, as estimativas são menos eficientes do que seria possível num modelo paramétrico.

De modo geral, esta menor eficiência não chegou a preocupar os usuários tradicionais do método. É possível provar (Efron, 1977) que a perda de eficiência não é grande, e sem dúvida é preferível sofrer um ligeiro aumento na variância dos estimadores do que precisar especificar a mal conhecida função $\mu_0(x)$ como parte do modelo. Se esta é a situação na maioria das aplicações bio-estatísticas, a demografia oferece condições algo diferentes. Poucas funções foram tão amplamente estudadas e codificadas em modelos quanto a variação da mortalidade com a idade numa população humana "normal", isto é, não-patológica. Com certeza, esta relação é conhecida em muito mais detalhe do que o impacto de variáveis como a educação ou a renda sobre o risco de morte em determinadas idades. Portanto, se o método dos riscos proporcionais é semi-paramétrico, dá para afirmar que é não-paramétrico na metade errada: não chega a ser uma estratégia muito racional especificar - com bastante rigidez - os aspectos mais obscuros do modelo para que - em compensação - seu componente mais ponderável possa ficar indefinido. Não é fácil formular um modelo que deixe livre a forma funcional da relação entre a

mortalidade e as variáveis independentes. O que é proposto aqui é mais simples, ou seja: já que aparentemente a propriedade semi-paramétrica não constitui nenhuma vantagem em análises demográficas da mortalidade, por que não usar uma parametrização da função de risco $\mu(x)$, aproveitando assim a maior eficiência deste tipo de estimativa?

Qual poderia ser esta parametrização? Não faltam modelos na literatura estatística que permitem a avaliação dos parâmetros de distribuições tradicionais, como a exponencial (Feigl & Zelen, 1965) ou a Weibull (Peto & Lee, 1973), em termos de uma regressão. Outra possibilidade é simplesmente parametrizar a família de riscos proporcionais, escolhendo alguma função específica para $\mu_0(x)$ (Hakkert, 1982). Entretanto, nem a primeira, nem a segunda alternativa parecem muito atraentes. A primeira porque, sabidamente, as distribuições estatísticas mais conhecidas são pouco adequadas à descrição da mortalidade humana. A segunda porque existem amplas evidências empíricas (ver Antonovsky, 1967) de que os diferenciais relativos da mortalidade humana não são constantes, mas variam conforme a idade, acentuando-se geralmente nas faixas etárias de 15 até 40 anos, e caindo marcadamente nas idades mais avançadas. Uma falta de proporcionalidade manifesta-se também - e talvez até com mais clareza - nos diferenciais da fecundidade, onde níveis diferentes geralmente são associados não apenas com taxas específicas mais ou menos altas em cada faixa etária, mas também com diferenças na tendência central e no desvio-padrão da distribuição etária. Evidentemente, esta crítica aplica-se não apenas ao variante paramétrico do modelo, mas igualmente a sua versão original. Principalmente em situações onde as idades dos indivíduos observados variam muito, este pode ser um problema mais grave do que a menor eficiência das estimativas semi-paramétricas, mencionada no início desta secção.

Fora estas duas críticas mais teóricas, existem algumas questões de conveniência. Seria conveniente, por exemplo, que determinados diferenciais que já fazem parte de práticas consagradas da análise demográfica, como o diferencial entre os sexos, recebessem um tratamento especial, não entrando, portanto, entre as variáveis independentes. Também seria conveniente que a variável resultante da regressão fosse um indicador demográfico comum, como por exemplo a esperança de vida, em vez do fator de proporcionalidade, cuja interpretação é algo menos óbvia. Sendo assim, facilitaria a identificação imediata dos coeficientes $\beta_1,\ldots,\beta_p$ com uma tábua de vida.

5. UMA ALTERNATIVA PARAMÉTRICA

Estuda-se atualmente a possibilidade de formular um método paramétrico baseado nas tábuas regionais modelo de Coale e Demeny (1983), ou outras famílias empíricas de um parâmetro, já consagradas pela prática demográfica. O método proposto atende a todas as críticas levantadas na secção anterior. Alguns outros aspectos da questão, que continuam sem respostas definitivas, serão comentados na próxima secção.

Na sua atual forma, o método parte do pressuposto de que a esperança de vida de cada indivíduo - ou, melhor, o nível, parametrizado pela esperança de vida feminina, da tábua de vida específica por sexo dentro do sistema modelo que caracteriza o indivíduo - pode ser decomposta em componentes baseados na série de variáveis $y_1,\ldots,y_p$:

$$e(0) = y = \beta_o + \beta_1 y_1 + \ldots\ldots + \beta_p y_p \qquad (8)$$

Por exemplo, para um homem, um resultado de 50 em (8) corresponderia ao nível 13 nas tábuas modelo de Coale e Demeny, ou seja, uma esperança de vida masculina de 47,08 anos no modelo Oeste, 46,70 nos modelos Norte e Leste, e 47,37 no modelo Sul. Para uma mulher, (8) expressa diretamente qual seria sua esperança de vida segundo o modelo.

Como na secção 3, o segundo passo é expressar a probabilidade dos acontecimentos verificados como função de $\beta_o,\ldots,\beta_p$. No caso, suponha que n indivíduos foram observados, com características $y_{1k},\ldots,y_{pk}$ $(k=1,\ldots,n)$. Além disso, suponha que o indivíduo k entrou na amostra com idade x_k e saiu com idade $x_k + t_k$, sendo que neste momento podia estar vivo $(\delta_k=0)$ ou morto $(\delta_k=1)$. Não é difícil avaliar, a partir destes pressupostos, que o logaritmo da função de verossimilhança L pode ser expressado da seguinte forma:

$$\log L = \sum_{k=1}^{n} \left((1-\delta_k) \log(\ell_{x_k+t_k}/\ell_{x_k}) + \delta_k \log(1-\ell_{x_k+t_k}/\ell_{x_k})\right) \qquad (9)$$

A função ℓ_x corresponde ao valor de e(0), definido em (8), e ao sexo de cada indivíduo. A tarefa agora se resume em encontrar os valores de $\beta_o,\ldots,\beta_p$ que maximizem a expressão (9). Possivelmente uma das razões por que este método até agora não foi usado nas análises da mortalidade diferencial seja que justamente este aspecto numérico, a minimização de (9) como função de $\beta_o,\ldots,\beta_p$, é tecnicamente um pouco mais complicado do que no caso de (7) ou dos modelos paramétricos mais comuns. O problema é que a distribuição ℓ_x não está dis-

ponível na forma de uma expressão analítica simples, mas encontra-se tabulada em intervalos de 5 (idade) e 2,5 (esperança de vida). anos. A maioria dos algoritmos computacionais mais usuais, como o clássico método de Newton-Raphson ou o mais recente de Powell e Fletcher (1963), exige que a função a ser maximizada seja diferenciável em cada ponto. Com isso, torna-se necessário encontrar uma interpolação relativamente sofisticada de ℓ_x. A solução pode estar na adoção de um dos algoritmos mais recentes que não requerem diferenciação, como o de Powell (1964) ou o método do simplex (Nelder & Mead, 1965). Estes minimizam qualquer função contínua, de modo que a interpolação das tábuas de vida modelo pode ser bem mais direta.

Nota-se ainda que é extremamente fácil incorporar mudanças nos valores de $y_1,\ldots,y_p$ de um mesmo indivíduo durante o período de observação, um problema que atrapalha a aplicação de métodos indiretos na análise de diferenciais. Afinal, estes exigem a classificação de cada mulher numa categoria única. No caso dos métodos de regressão, tanto o dos riscos proporcionais quanto o paramétrico proposto aqui, esta complicação pode ser resolvida simplesmente pela divisão da trajetória do indivíduo em duas etapas, uma com os valores anteriores de $y_1,\ldots y_p$, e uma com os valores novos. Do ponto de vista formal é como se se tratasse de duas pessoas distintas.

6. ALGUNS ASPECTOS NÃO RESOLVIDOS

Além das críticas levantadas na secção 4, existem algumas complicações, tanto do modelo dos riscos proporcionais quanto do método traçado na secção anterior, que continuarão exigindo atenção. Como já se notou antes, a fórmula (5) não inclui um termo E expressivo do erro aleatório, como no modelo de regressão comum. O mesmo acontece com (8). A ausência de um termo de erro implica no suposto de que toda a heterogeneidade existente nos níveis da mortalidade é captada pelo modelo, ou seja, não há nenhum outro fator, não incorporado no modelo, que possa acrescentar um componente adicional de variação entre os indivíduos. Como foi demonstrado por Vaupel et al. (1979), uma heterogeneidade não explicada deste tipo tende a distorcer o padrão etário observado da função $\mu_o(x)$, deprimindo-a nas idades mais avançadas. A causa é a tendência dos indivíduos menos resistentes a morrerem mais cedo, de modo que são apenas os mais fortes que alcançam estas idades mais avançadas. A conseqüência pode ser um certo viés nas tábuas de vida estimadas pelo método dos riscos proporcionais. Vaupel et al. (1981) tentaram resolver este problema, estimando o grau de heterogeneidade residual através de uma dis-

tribuição gama. Heckman e Singer (1982), entretanto, demonstraram que estimativas paramétricas deste tipo são pouco robustas, isto é, dependem muito da família específica de distribuições escolhida. Eles propõem um método não paramétrico, mas este exige que a forma de $\mu_0(x)$ seja conhecida. Conforme mostram Trussell e Richards (1983), o método de Heckman e Singer resulta em estimativas que dependem fortemente do tipo de função escolhido para $\mu_0(x)$. A conclusão é que, dentro da formulação do modelo dos riscos proporcionais, não há solução: no mínimo uma das funções, $\mu_0(x)$ ou a distribuição de E, deve ser explicitada e o resultado é fortemente afetado por esta opção. Na formulação paramétrica da secção 5, o problema não se apresenta de forma tão saliente, já que o padrão etário das tábuas de vida resultantes faz parte da estrutura do próprio método, ou seja, a família específica de tábuas de vida modelo que for adotada. Entretanto, é possível que formas de heterogeneidade não explicadas pelos fatores $y_1,\ldots,y_p$ introduzam um certo viés nas estimativas dos parâmetros $\beta_0,\ldots,\beta_p$. É muito cedo para avaliar a gravidade deste problema, mas sem dúvida é uma questão a ser investigada.

Outro problema diz respeito às interações entre a idade x e os fatores $y_1,\ldots,y_p$. Como mostra Chackiel (1982), o impacto dos diversos fatores sobre a mortalidade infantil e de crianças depende, até certo ponto, da idade do indivíduo. Especificamente, Chackiel encontra uma relação estreita entre a mortalidade das crianças mais jovens e os determinantes biológicos, relação que tende a se atenuar com o decorrer dos anos, quando é gradualmente substituída por um papel mais preponderante dos determinantes sócio-econômicos. À primeira vista, esta constatação parece invalidar tanto a especificação (5) quanto (8). Entretanto, nada impede que $y_1,\ldots,y_p$ sejam definidos como funções das variáveis substantivas a serem analisadas e da idade x. Por exemplo, se se desconfia que o efeito do peso ao nascer diminui exponencialmente com a idade, é possível definir uma variável independente $y_j = P \exp(-ax)$ para expressar esta relação. A dificuldade, neste caso, estaria na determinação de um valor mais ou menos realístico para a, enquanto também seria necessário dividir a trajetória de cada indivíduo em intervalos de um ano para acomodar as mudanças de y_j com o tempo.

Finalmente, existe a questão qual modelo expressa melhor o impacto das variáveis independentes sobre a mortalidade, (5) ou (8). Este, evidentemente, é um ponto a ser esclarecido empiricamente. Entretanto, em pelo menos um aspecto a especificação (5) parece algo mais realística. Nota-se que

na formulação (8) qualquer mudança em uma das variáveis $y_1,\ldots,y_p$ sempre exerce o mesmo efeito sobre a esperança de vida, seja qual esta for. A formulação (5), ao contrário, contém uma propriedade de saturação, já que uma mudança relativa de $\mu(x)$ tende a elevar a esperança de vida menos quando esta já está alta. Numa formulação algo mais sofisticada de (8), entretanto, seria possível corrigir este defeito pela inclusão de uma função logística ou outra, com características assintóticas semelhantes, para converter y em e(0).

7. MÉTODOS DE REGRESSÃO PARA ESTIMATIVAS INDIRETAS

No final deste artigo, cabe fazer alguma menção dos desenvolvimentos metodológicos recentes que visam aplicar modelos de regressão a estimativas obtidas por meios indiretos. Os avanços nesta área não dependem analiticamente da formulação dos modelos comentados anteriormente, mas certamente foram inspirados por eles.

O trabalho mais conhecido na área é o de Trussell e Preston (1982), que distinguiram uma variedade de metodologias a serem adotadas, dependendo do grau de detalhe contido nos dados: apenas os números de filhos nascidos vivos e sobreviventes por idade da mulher, ou também todas ou algumas datas de nascimento. Será comentado apenas o caso mais comum, onde a informação disponível se limita ao número de nascidos vivos e sobreviventes. Neste caso, os autores fazem um pressuposto parecido com (4), ou seja:

$$1 - \ell_x = \lambda\ (1 - \ell^o_x) \qquad (10)$$

Para valores de ℓ_x e ℓ^o_x próximos de 1, como geralmente é o caso na análise da mortalidade infantil e de crianças, (4) e (10) são aproximadamente equivalentes. Trussell e Preston escolhem uma abordagem paramétrica. Baseando-se nas tábuas de vida modelo de Coale e Demeny, calculam qual seria a proporção de filhos mortos D^o_i na faixa etária i de mulheres se a tábua de vida fosse ℓ^o_i, uma função pré-definida que pode ser, por exemplo, a tábua de vida do conjunto da população. A divisão da proporção $D_i(y)$ efetivamente verificada na categoria de mulheres com as características $y_1,\ldots,y_p$ pelo valor teórico D^o_i leva a λ. Este parâmetro, finalmente, pode ser especificado em termos de vários tipos de regressão; no trabalho citado, Trussell e Preston usam a regressão linear comum e a regressão por tobits (ver Tobin, 1958). Uma alternativa seria usar (5), ou seja, uma regressão linear de log $D_i(y)/D^o_i$ em termos de $y_1,\ldots,y_p$; neste caso, a possibilidade de que $D_i(y)=0$ requer algumas precauções. Em todos os casos, é

recomendado usar uma ponderação de cada observação pelo número de filhos nascidos vivos que serve como base para a avaliação de $D_i(y)$ e D_i^o.

Existem algumas variações sobre o tema. A avaliação de D_i^o, por exemplo, pode ser feita com base no esquema de fecundidade vigente no conjunto da população, ou especificamente com base na fecundidade verificada dentro do grupo com as características $y_1,\ldots,y_p$. Teoricamente, a segunda alternativa seria mais correta, mas na prática existe o risco de estimativas instáveis por causa do reduzido número de mulheres em cada categoria. Outra opção diz respeito ao uso de variáveis categóricas ou contínuas para $y_1,\ldots,y_p$. Quando se usa o esquema específico de fecundidade de cada grupo de mulheres, ou quando o procedimento de regressão adotado tem $\log D_i(y)/D_i^o$ como variável dependente, é inevitável usar variáveis independentes categorizadas para manter um número suficiente de mulheres em cada ponto de observação e evitar muitas ocorrências de $D_i(y)=0$. Noutras circunstâncias, entretanto, é possível aplicar o método a nível de cada mulher, usando variáveis contínuas (Schultz, 1979).

Uma contribuição mais recente, de Choe (1983), segue um raciocínio muito semelhante, ficando, entretanto, mais próximo do método dos riscos proporcionais em dois aspectos. Primeiro, em vez de usar a expressão (10) proposta por Trussell e Preston, Choe parte de (4) ou, mais precisamente, da seguinte fórmula equivalente:

$$\log(-\log(\ell_x)) = \log(-\log(\ell_x^o)) + \beta_1 y_1 + \ldots\ldots + \beta_p y_p \quad (11)$$

onde a especificação (5) do modelo de regressão já foi incorporada. A partir desta equação, ela demonstra que uma relação semelhante existe entre $S_i(y)$, a proporção de filhos sobreviventes das mulheres na faixa etária i com as características $y_1,\ldots,y_p$, e S_i^o, a proporção que corresponderia à tábua de vida ℓ_x^o, ou seja:

$$\log(-\log(S_i(y))) = \log(-\log(S_i^o)) + \beta_1 y_1 + \ldots\ldots + \beta_p y_p \quad (12)$$

Para que esta relação seja válida, entretanto, é necessário que todos os esquemas de fecundidade sejam iguais. Além disso, geralmente será necessário usar variáveis independentes categorizadas, para evitar que $S_i(y)$ alcance os valores extremos de 0 ou 1.

A segunda diferença entre a abordagem de Choe e a de Trussell e Preston é que a expressão (12) permite um procedi

mento semi-paramétrico, conforme o modelo original dos riscos proporcionais. Afinal, o termo $\log(-\log(S_i))$ depende apenas de i, de modo que não é preciso fazer qualquer pressuposto sobre este termo. Seu valor pode ser determinado como parte do procedimento de regressão comum, seja limitando este procedimento a uma faixa etária das mulheres cada vez, seja pela divisão de $\log(-\log(S_i))$ numa série de termos "dummy" para expressar o efeito da faixa etária da mãe. Apesar do pressuposto algo restritivo de esquemas uniformes de fecundidade para todas as mulheres, o método de Choe parece bastante atraente em circunstâncias onde as variáveis explicativas da mortalidade possuem uma categorização mais ou menos natural (sexo, classe social, situação urbana ou rural, etc.) e onde a priori o padrão etário subjacente da mortalidade é pouco conhecido.

REFERÊNCIAS BIBLIOGRÁFICAS

ANTONOVSKY, Aaron. 1967. Social class, life expectancy and overall mortality. Milbank Memorial Fund Quarterly, 45: :31-73.

AVERY, Roger C. & HAKKERT, Ralph. 1981. Comment on "The random variation in rates based on total enumeration of events" by J. R. Udry, C. Teddlie and C. H. Suchindran. Population Studies, London, 36 (3): 467-71.

BAKER, R. J. & NELDER, J. A. 1978. The GLIM system manual - release 3. Numerical Algorithms Group, Oxford.

BEHM ROSAS, Hugo et alii. 1976-80. Mortalidad en los primeros años de vida en países de la America Latina. San José, CELADE, Série A 1024-32, 1036-37.

BISHOP, Yvonne M. et alii. 1975. Discrete multivariate analysis: theory and practice. Cambridge MA, MIT Press.

BRESLOW, Norman E. 1975. Analysis of survival data under a proportional hazards model. International Statistics Review, 43: 45-57.

CHACKIEL, Juan. 1982. Factores que afectan a la mortalidad en la niñez. Notas de Población, 10 (28): 43-85.

CHOE, Minja Kim. 1983. The indirect proportional hazards model: an adaptation of the proportional hazards model to the indirect method of estimating infant and childhood mortality. Trabalho apresentado na Reunião da PAA, Pittsburgh.

COALE, Ansley J. & DEMENY, Paul. 1983. Regional model life tables and stable populations. New York, Academic Press.

COX, David R. 1972. Regression models and life tables. Journal of the Royal Statistical Society, Series B 34 (2): 187-220.

COX, David R. 1975. Partial likelihood. Biometrika, 62: 269--76.

EFRON, B. 1977. The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Society, 72: 557-65.

ELANDT-JOHNSON, R. C. & JOHNSON, N. L. 1980. Survival models and data analysis. New York, Wiley & Sons.

FEIGL, P. & ZELEN, M. 1965. Estimation of exponential survival probabilities with concomitant information. Biometrics, 21: 826-38.

FOX, A. J. 1979. Prospects for change in differential mortality. In: UN/WHO. Proceedings of the Meeting on Socio-economic Determinants and Consequences of Mortality. México, Colegio de México.

FRENZEN, P. D. & HOGAN, D. P. 1982. The impact of class, education, and health care on infant mortality in a developing society: the case of rural Thailand. Demography, 19 (3): 391-408.

GOODMAN, Leo A. 1953. Ecological regressions and behavior of individuals. American Sociological Review, 18: 663-4.

GORTMAKER, Steven L. 1979. Poverty and infant mortality in the United States. American Sociological Review, 44: 280-97.

HAKKERT, Ralph. 1982. Mortalidade proporcional com padrão etário dado: avaliação de parâmetros e adequação do modelo. Trabalho apresentado na 3a. Reunião da ABEP, Vitória.

HAKKERT, Ralph. 1984. Trends and differentials of mortality in Brazil, 1950-1975. Tese de Doutorado, Universidade de Cornell, Ithaca NY.

HECKMAN, James J. & SINGER, Burton. 1982. Population heterogeneity in demographic models. In: LAND, Kenneth & ROGERS, Andrei (eds.). Multidimensional mathematical demography. New York, Academic Press: 567-99.

HOBCRAFT, John et alii. 1982. Socio-economic factors in infant and child mortality: cross-national comparisons. Trabalho apresentado na Reunião da PAA, San Diego.

KALBFLEISCH, John D. & PRENTICE, Ross L. 1980. The statistical analysis of failure time data. New York, Wiley & Sons.

KEYFITZ, Nathan & LITTMAN, G. 1980. Mortality in a heterogeneous population. Population Studies, London, 33 (2): 333-43.

KITAGAWA, Evelyn M. & HAUSER, Philip M. 1973. Differential mortality in the United States. Cambridge MA, Harvard University Press.

LAIRD, Nan & OLIVIER, Donald. 1981. Covariance analysis of censored survival data using log-linear analysis techniques. Journal of the American Statistical Association, 76: 231-40.

LAWLESS, Jerald F. 1982. Statistical models and methods for life time data. New York, John Wiley & Sons.

MANTON, Kenneth G. et alii. 1981. Methods for comparing the mortality experience of heterogeneous populations. Demography, 18 (3): 389-410.

MARTIN, Linda G. et alii. 1983. Covariates of child mortality in the Philippines, Indonesia and Pakistan: an analysis based on hazard models. Population Studies, London, 37 (3): 417-32.

MENKEN, Jane et alii. 1981. Proportional hazards life table models: an illustrative analysis of socio-demographic influences on marriage dissolution in the United States. Demography, 18 (2): 181-200.

MEYER, Paul L. 1969. Probabilidade: aplicações à estatísti-

ca. Rio de Janeiro, Livros Técnicos e Científicos Ed.

MICHAEL, Robert T. & TUMA, Nancy B. 1983. Entry into marriage and parenthood by young adults. Trabalho apresentado na Reunião Anual da PAA, Pittsburgh.

OLIVIER, Donald & NEFF, Raymond. LOGLIN 1.0: user's guide. Cambridge MA, Harvard University Health Science Computing Facility.

PETO, R. & LEE, P. 1973. Weibull distributions for continuous carcinogenesis experiments. Biometrics, 29: 457-70.

ROBINSON, W. W. 1960. Ecological correlation and the behavior of individuals. American Sociological Review, 15: 351--7.

RODGERS, Gerry B. 1979. Income and inequality as determinants of mortality: an international cross-section analysis. Population Studies, London, 33 (2): 343-51.

SANDEFUR, Gary & SCOTT, Wilbur. 1981. A dynamic analysis of migration: an assessment of the effects of age, family and career variables. Demography, 18 (3): 355-368.

SANTOS, Jair Lício F. et alii. 1976. A mortalidade no Brasil em 1970. Trabalho apresentado no Simpósio sobre o Progresso da Pesquisa Demográfica no Brasil.

SCHULTZ, T. Paul. 1979. Interpretation of relations among mortality, economics of the household, and the health environment. In: UN/WHO. Proceedings of the Meeting on Socioeconomic determinants and Consequences of Mortality. México, Colegio de México.

SHEPARD, D. S. & ZECKHAUSER, R. J. 1975. The assessment of programs to prolong life, recognizing their interaction with risk factors. Cambridge MA, Kennedy School of Government, Discussion Paper 32-D, Harvard University.

SHEPARD, D. S. & ZECKHAUSER, R. J. 1977. Heterogeneity among patients as a risk factor in surgical decision-making. In: BUNKER, J. P., et alii (ed.). Costs, risks, and benefits of surgery. New York, Oxford University Press.

SIMÕES, Celso C. da S. 1981. O quadro da mortalidade por classes de renda: um estudo dos diferenciais nas regiões metropolitanas (núcleo e periferia). Tese de Mestrado,

UFRJ, Rio de Janeiro.

SINGER, Paul I. et alii. 1978. Prevenir e curar: o controle social através dos serviços de saúde. Rio de Janeiro, Forense Universitária.

TOBIN, James. 1958. Estimates of relationships for limited dependent variables. Econometrica, 26: 24-36.

TRUSSELL, T. James & BLOOM, D. E. 1983. Estimating the co--variates of age at marriage and age at first birth. Population Studies, London, 37 (3): 403-16.

TRUSSELL, T. James & HAMMERSLOUGH, Charles. 1983. A hazards--model analysis of the covariates of infant and child mortality in Sri Lanka. Demography, 20 (1): 1-26.

TRUSSELL, T. James & PRESTON, Samuel H. 1982. Estimating the covariates of childhood mortality. Health Policy and Education, Amsterdam, 3: 1-36.

TRUSSELL, T. James & RICHARDS, Toni. 1983. Correcting for unobserved heterogeneity in hazard models: an application of the Heckman-Singer procedure to demographic data. Trabalho apresentado na Reunião da PAA, Pittsburgh.

TUMA, Nancy B. 1982. Non-parametric and partially parametric approaches to event history analysis. In: LEINHARDT, S. (ed.). Sociological methodology. San Francisco, Jossey--Bass.

TUMA, Nancy B. et alii. 1979. Dynamic analysis of event his tories. American Journal of Sociology, 84 (4): 820-54.

VAUPEL, James W. et alii. 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography, 16 (3): 439-54.

VAUPEL, James W. & YASHIN, Anatoli I. 1983. The deviant dynamics of death in heterogeneous populations. Laxenburg, Austria, IIASA Research Report 83-1.

WYMAN, Kathie E. 1981. Application of Cox model and other life table techniques to the study of work-fertility relationship. Trabalho apresentado na Reunião da PAA, Washington DC.

ABSTRACT - Some recent techniques are discussed which facilitate the analysis of differential mortality according to several simultaneous criteria, in situations where small samples impede cross-tabulation by all relevant dimensions. All of these techniques involve the use of regression models. Although some technical details are inevitable, the primary objective is to provide an introduction to the subject for the non-specialists, with an ample bibliography of the current theoretical and applied literature in the area. The logic behind Cox's proportional hazards method, the most common in present applications, is explained, and its advantages and disadvantages are pointed out. More appropriate alternatives for the analysis of mortality in human populations are indicated. Finally, some comments are made on the problem of unexplained heterogeneity and on extensions to indirect estimation, which were inspired by Cox's method.