Métodos Estatísticos na Análise de Mudanças Ocupacionais*

PAULO VIEIRA DA CUNHA**

Introdução

A análise estatística de processos de mobilidade ocupacional baseia-se, costumeiramente, em tabelas bidimensionais referentes à situação passada e presente de uma determinada população(1). De certo, este tipo de comparação é útil, porém, tal utilidade seria maior se fosse possível considerar estruturas tabulares mais complexas. Afinal, o debate teórico atual tem privilegiado discussões sobre o *processo* de mobilidade, entendido como uma relação entre um conjunto de características pessoais e estruturais e os resultados ocupacionais observados. Isto posto, é importante recordar que o processo em questão é necessariamente heurístico; em parte, porque os dados empíricos disponíveis referem-se a fontes amostrais, mas, também, porque, ao considerar o processo de mudança ocupacional, no nível individual, uma especificação estatisticamente adequada não poderá prescindir de um elemento de chance. Decorre disso a necessidade de formulação do exercício empírico no contexto da teoria formal de testes e hipóteses, estendida a tabelas de contingências multidimensionais.

Estas tabelas apresentam problemas especiais de análise e interpretação. Neste trabalho, pretendemos discutir tais problemas, relacionando-os com o que nos inte-

* Este trabalho apóia-se no segundo capítulo de nossa tese de doutoramento "Structures of Production and Employment. Occupation Change in Monterrey, Mexico: 1940-1950" apresentada à Universidade da Califórnia (Berkeley), em 1980. Somos particularmente gratos ao professor Albert Fishlow, presidente da comissão de tese, pelas suas críticas, sugestões e estímulo. Agradecemos, ainda, ao professor Mark Nerlove, da Northwestern University, por possibilitar acesso a seu trabalho sobre modelos loglineares. Bolsas da FAPESP, do Convênio FORD/CEBRAP e do Programa ECIEL contribuíram, em diferentes períodos, para a execução deste trabalho.

** Do IPEA/INPES e do Departamento de Economia da UFRJ.

(1) Ver, por exemplo, o estudo clássico de Palmer (1954) e o trabalho de Pastore (1979) sobre os dados da PNAD. Goodman (1969) apresenta uma excelente crítica metodológica destes métodos.

ressa de imediato, mas, igualmente, reportando-nos à vasta e crescente literatura disponível sobre o método. Como observou Fienberg, ao ter de analisar classificações em tabelas multidimensionais, a maioria dos pesquisadores encararia o problema desagregando a tabela em vários subconjuntos dimensionais, ou seja, "examinaria as variáveis categóricas duas de cada vez" Entretanto, uma abordagem deste tipo: "(a) faz com que se confunda a relação marginal entre duas variáveis categóricas com a relação existente, quando outras variáveis se encontram presentes, (b) não permite que se proceda ao exame simultâneo das múltiplas relações binárias, (c) ignora a possibilidade de existência de interações tripartides e de ordem superior entre as variáveis."[2].

Considere-se, por exemplo, a interação entre mudança ocupacional (*O*), idade (*I*) e formação educacional (*E*). Qualquer inferência baseada na tabela *O* x *I*, isoladamente, será enganosa, visto que é provável que *I* e *E* interajam entre si. Além disto, do ponto de vista analítico, podemos estar interessados não somente nas interações separadas, *O* x *I* e *I* x *E*, mas em como a educação agindo por intermédio da idade afeta a mudança ocupacional (ou, alternativamente, em como a mudança ocupacional, filtrada pela idade, afeta a formação educacional). É claro que é possível repartir qualquer tabela multidimensional, dando-lhe duas dimensões. Entretanto, como Simpson (1951) demonstrou, tal procedimento pode facilmente provocar incorreção, como é visto nas interações entre *O* x *I* x *E* apresentadas na tabela abaixo.

Mobilidade Ocupacional	Educação					
	Trabalhador Jovem		Trabalhador mais Velho		Todos os Trabalhadores	
	Nenhuma	Alguma	Nenhuma	Alguma	Nenhuma	Alguma
Móveis	0	20	20	0	20	20
Fixos	20	0	0	20	20	20

O exemplo dado é, sem dúvida, um caso extremo e só um analista desatento não veria que o efeito da formação educacional na perspectiva profissional depende basicamente da idade. A lição que tiramos daqui é que, caso existam associações envolvendo mais do que duas variáveis, estas podem se tornar totalmente incertas quando agregamos a tabela. Também é possível, mediante a agregação indevida, produzir uma tabela que, aparentemente, apresente uma interação num contexto em que as duas variáveis sejam, por qualquer combinação das variáveis restantes, independentes uma da outra.

O recente desenvolvimento da teoria estatística sobre modelos loglineares aplicados à análise de tabelas de contingência multidimensionais oferece soluções para as principais limitações anteriores assinaladas[3]. Entretanto, interessa-nos, aqui, um tipo particular de modelo loglinear; neste modelo é possível diferenciar dois subconjuntos de variáveis, de um lado as variáveis explicativas e, de outro, aquelas (em nosso caso, é, de fato, apenas uma) endogenamente determinadas pelas primeiras. Ademais, em tal modelo, pelo menos algumas variáveis explicativas são contínuas. Esta característi-

(2) Fienberg (1977, p. 1).

(3) Fienberg (1977, p. 4-2) apresenta um rápido apanhado histórico sobre a situação anterior ao desenvolvimento de modelos loglineares. Seu livro apresenta de maneira concisa e clara o método básico e as propriedades estatísticas dos modelos loglineares. Uma apresentação teórica mais completa pode ser encontrada em Bishop, Fienberg e Holland (1975) e nas contribuições de Nerlove e Press (1973), (1976) e (1977), Haberman (1974) e Goodman (1970).

ca conduz à adoção de uma transformação logística estabelecida a partir da razão de probabilidades entre as categorias da variável dependente. Mas, de fato, os modelos logísticos são, como iremos demonstrar, expressões particulares do modelo loglinear geral. Por esse motivo, a apresentação do método está organizada da seguinte maneira: em primeiro lugar, introduzimos a anotação loglinear para tabelas simples de dupla entrada, em seguida, aplicamos este procedimento ao caso geral de tabelas mais complexas. Discutimos, então, os problemas de inferência estatística em modelos loglineares e, finalmente, mostramos como a formulação logística pode derivar, em casos especiais, do modelo loglinear geral.

Existem, entretanto, duas questões que devem ser tratadas previamente, às quais faremos breve referência: o tratamento dado à variável tempo na análise de mobilidade (isto é, a definição de períodos) e o porquê da escolha (algo raro em trabalhos de economia) de modelos loglineares e não logísticos[4].

1. Período Biográfico e Escolha de Modelos

O fator tempo, embora difícil de ser tratado, é um problema fundamental na pesquisa sobre mobilidade. Sem dúvida, ao se analisar padrões de mobilidade, seria melhor considerar toda a *carreira* de um indivíduo como o elemento básico de observação. Porém, optamos por uma definição de mudanças discretas ao longo do tempo; vale dizer, por uma paralisação da carreira profissional.

Em parte, esta opção decorre de nossa preocupação específica, já que interessa-nos, sobretudo, relacionar os resultados ocupacionais às transformações na estrutura de produção. Para tanto, devemos fixar períodos comparativos que tenham por base a evolução da produção e sejam, portanto, iguais para todos os trabalhadores incluídos na amostra. Existe também um outro motivo para a escolha: os modelos baseados em seqüências contínuas de ocupações são empiricamente impraticáveis ou conceptualmente simplistas. Caso o intervalo entre ocupações seja tratado como uma variável quase-contínua, o número de observações em cada célula das tabelas de mobilidade correspondentes a cada intervalo torna-se pequeno demais para a implementação de um modelo estatístico[5]. Nesta situação, uma alternativa viável seria adotar um modelo específico para a seqüência temporal de probabilidades de mudança ocupacional, tal como uma cadeia de Markov[6].

São muitos e recentes os exemplos deste tipo de trabalho, mas, na maioria das vezes, os esforços ainda carecem de con-

(4) McFadden (1976a) resenha as obras mais importantes. Boskin (1973) e Schmidt e Strauss (1975) trabalharam com modelos de escolha ocupacional que podem ser, à primeira vista, considerados (erroneamente) como tendo características similares às do nosso problema específico. Estas análises não exploram o processo de mudança ocupacional. Em vez disto, a partir de dados de "cross-section" tentam prever as probabilidades de um trabalhador estar em uma dada ocupação. O objetivo explícito destes exercícios é "testar" uma teoria e não explorar a estrutura do processo de mobilidade.

(5) A esse respeito, ver o artigo clássico de Duncan (1966).

(6) Ao analisarem os dados de pesquisa de Monterrey, Balán, Browning e Jelin (1973) apresentam uma outra alternativa. Seu trabalho baseia-se, principalmente, nos modelos longitudinais do tipo desenvolvido por Blau e Duncan (1967). Os modelos examinados no capítulo 10 do livro supracitado são alimentados por observações sobre a primeira ocupação do entrevistado e, dependendo do grupo etário, a ocupação nas idades de 25, 35, 45 e 55 anos. Note que apesar de serem possíveis comparações entre os grupos etários, a análise não leva em consideração (pelo menos não explicitamente) a ligação existente entre o padrão variável das demandas de mão-de-obra e o processo de mudança ocupacional. Ver Sewell e Hauser (1975) para um trabalho posterior sobre a mesma família de modelos.

teúdo substantivo[7]. Falta aos modelos uma especificação mais complexa do processo de mobilidade — por uma razão muito simples. Sua operacionalidade está condicionada às hipóteses do modelo probabilístico, como no exemplo de um processo markovniano. O esforço analítico ocorre no sentido de adaptar o processo de mobilidade ao modelo probabilístico conhecido, não na direção de elaborar um modelo para o processo. Esta última alternativa poderia, de certo, proporcionar modelos com propriedades desconhecidas e, portanto, sem aplicabilidade empírica[8].

Por outro lado, quando os modelos são modificados para levar em consideração a instabilidade e heterogeneidade das matrizes de transição (ver Spilerman (1972)), as técnicas de estimação sugeridas aplicam-se unicamente a probabilidades de uma só transição. Ademais, elas estão baseadas (ou deveriam basear-se) em modelos loglineares ou logísticos do tipo que adotaremos a seguir[9]. As tentativas que forem feitas para montar modelos Markov estacionários e homogêneos, com matrizes de transição formadas pela média obtida entre matrizes não-estacionárias e heterogêneas construídas a partir de uma seqüência discreta de transições, cada uma unindo apenas um período (como sugerido por McFarland (1970)), certamente falharão. Afora o problema de estimação (que seria grande), deve-se reconhecer que a mobilidade ocupacional envolve mudanças que são fundamentalmente determinadas pela estrutura das demandas de mão-de-obra. A distribuição limite do modelo (ergótico) de Markov, isto é, a definição de matriz de transições do estado "estacionário" ou "fixo" não reflete de forma alguma esta situação. Aqui, as probabilidades de transição determinam *per se* a distribuição ocupacional, contrariamente à noção de que a disponibilidade de empregos em diferentes grupos ocupacionais influencia a mobilidade para essas ocupações. Assim, no modelo Mar-

(7) Para uma excelente resenha, consulte-se o trabalho de Stewman (1976) e, para uma apresentação formal do "estado da arte," o trabalho de Singer e Spilerman (1976). Um trabalho pioneiro sobre processos semi-Markov é o de Blumen, Kogan e MacCarthy (1955), ao qual Prais (1955) fez importantes contribuições, bem como Hodge (1966), McFarland (1970), Spilerman (1972), Theil (1972) — capítulo 5 — e Sorensen (1975). Uma revisão e a sugestão de um modelo podem ser encontrados em Vieira da Cunha (1975). Modelos modificados de trabalhadores móveis estáveis foram aplicados ao processo de mobilidade entre categorias de renda por MacCall (1973) e Shorrocks (1976). Lillard e Willis (1978), num artigo de envergadura, criticam esta abordagem e sugerem um novo modelo para dinâmicas de renda que não são, entretanto, markovnianas.

(8) Após pesquisar as várias aplicações de processos de cadeias de Markov, Ginsberg chegou à seguinte conclusão: "Na área da mobilidade, os modelos markovnianos de mobilidade da mão-de-obra industrial não estão relacionados com os estudos sobre a estrutura industrial e o crescimento econômico. Também não há qualquer relação entre modelos markovnianos de mobilidade entre gerações e estudos sobre oportunidade econômica e educacional, estilo de vida, organização de classes etc... Elaboramos análise matemática da estrutura estocástica do processo, ignorando explicitamente seus determinantes" Ginsberg (1972, p. 66).

(9) Diz-se que um processo é não-estacionário quando, ao se analisar a mudança ocupacional, a matriz de probabilidade de transição varia com o tempo. A heterogeneidade refere-se ao fato de que a matriz não é a mesma para todos os indivíduos, mas difere em resposta a um vetor conhecido de características individuais. Em artigo publicado em 1972, Spilerman defende o uso de modelos lineares em probabilidades para avaliar as heterogeneidades nas células de uma matriz de transição. As aplicações do modelo linear em probabilidades e com variáveis exógenas contínuas levam a estimativas distorcidas e ineficientes das probabilidades, resultado este já apontado por Theil (1970) e elaborado por Kon (1976, p. 19-22). As condições para equivalência entre os resultados dos modelos linear e linear em probabilidades são formalmente demonstradas em MacRae (1977); ver também Wise (1975, p. 921).

kov, é a mobilidade que determina a estrutura e não a estrutura que determina a mobilidade[10].

Voltando agora à questão de modelos loglineares *versus* modelos logísticos, deve-se reconhecer de imediato que *não é a* técnica de estimação empregada que os separa. Apesar das grandes diferenças conceptuais na formulação do problema probabilístico, a técnica atual de estimação é a mesma, seja qual for a formalização teórica que venha a ser escolhida. Parafraseando McFadden, poderíamos dizer que o experimento padrão (para um grupo-população-de trabalhadores em uma dada ocupação e ano) consiste na seleção de uma amostra de observações sobre mudanças ocupacionais que definem, simultaneamente, as características individuais e os destinos ocupacionais dos trabalhadores escolhidos. Em outras palavras, as mudanças são sensíveis às variações nos atributos das ocupações de destino e às diferenças, entre indivíduos, em suas características pessoais. Supondo que o número de ocupações seja limitado (isto é, discreto), podemos postular que os resultados obtidos do experimento provêm de uma distribuição multinominal, com probabilidades de seleção condicionadas à ocupação de origem, aos valores observados nas características individuais e aos atributos dos destinos ocupacionais[11].

A diferença entre os dois modelos está no fato de que é possível derivar a especificação logística da teoria neoclássica de escolha individual. Neste caso, seria necessário introduzir a hipótese adicional e arbitrária de que os termos de perturbação randômica nas funções individuais de utilidade correspondentes ao exercício de escolha na forma logística são independentes e identicamente distribuídos segundo uma classificação exponencial recíproca ou Weibull[12]. Os modelos loglineares, por outro lado, são simples descrições matemáticas de tabelas de probabilidades, que mediante algumas suposições sobre a relação existente entre as variáveis — suposições estas expressas em termos de distribuições conjuntas de probabilidade para cada observação na amostra — podem-se transformar em modelos logísticos. Assim, o uso da especificação logística não implica qualquer hipótese restritiva sobre a distribuição de diferenças individuais não observadas.

Já que não pretendemos sobrecarregar nosso modelo empírico com o peso de uma teoria axiomática de escolha, preferimos a segunda alternativa. Mesmo que decidamos ignorar a crítica metodológica que possa ser feita aos conceitos de maximização da utilidade e escolha subjetiva ainda assim não podemos adotar a teoria, pois a estrutura teórica formal necessária simplesmente inexiste. Neste caso, far-se-ia necessária uma teoria operacionalmente válida de maximi-

(10) Ver Singer e Spilerman (1974).

(11) McFadden (1976b, p. 511). Neste particular, a abordagem difere da técnica de **análise discriminante** "que postula que os valores observados das características individuais e dos atributos das alternativas são amostras colhidas de distribuições posteriores ao evento condicionadas nas escolhas efetivamente observadas **(ibid.)**. O emprego da análise discriminante para estimar modelos logísticos levará, geralmente, a resultados distorcidos — **ibid.** (p. 517-521) e Press & Wilson (1978, p. 801).

(12) Ver McFadden (1972, Lemma 1, p. 8-9). Não há razão "econômica" para esta escolha: "Infelizmente, esta especificação precisa ser feita por razões de cálculo, já que ela é a única distribuição de probabilidades que conduz a uma função de verossimilhança de alguma simplicidade" (Kohn, Manski e Mundel, 1976, p. 395). Dentro do contexto da teoria neoclássica da maximização de utilidade, a principal desvantagem da forma funcional, estabelecendo as bases do "logit" condicional, é a propriedade denominada "independência das alternativas irrelevantes" De acordo com McFadden (1976a. p. 369), "este axioma estabelece aproximadamente que as chances em uma escolha binária permanecerão as mesmas para as alternativas em consideração quando outras alternativas adicionais estiverem disponíveis". Esta propriedade tem sido justificadamente questionada. Ver, por exemplo, Hausman e Wise (1978).

zação de utilidade com escolha ocupacional ótima, em um contexto dinâmico. Não temos, certamente, capacidade ou inclinação para desenvolver esta última e, apesar de haver algumas tentativas neste sentido, as restrições acima mencionadas tornam estes modelos pouco aconselháveis para aplicações empíricas[13].

Adicionalmente, na teoria da escolha da literatura econométrica, são as características das *escolhas* que compõem a função de probabilidade, enquanto os parâmetros referem-se às características do tomador de decisão. Da forma como utilizamos a transformação logística, os argumentos de função referem-se às características dos trabalhadores — as características das escolhas (destino ocupacional) refletem-se nos parâmetros da função. Mesmo se quiséssemos, não poderíamos estimar um modelo de escolha ocupacional. Afora a informação empregada para classificar as ocupações, não existem dados na nossa amostra sobre as características das escolhas em si[14]. Não podemos saber quanto um trabalhador estaria ganhando, caso estivesse transferindo-se para esta ou aquela ocupação, ou durante quanto tempo teria de receber treinamento para tal etc. Também não temos interesse nestes tipos de exercício, pois eles distanciam-nos do que consideramos extremamente importante: a estrutura de produção e demandas de mão-de-obra.

2. O Modelo Loglinear[15]

Considere-se uma amostra de N indivíduos $(n=1,\ldots,N)$ agrupados nas quatro células de uma tabela de mobilidade ocupacional unindo dois períodos e duas ocupações. Por construção, teríamos classificado o total da amostra em dois grupos mutuamente exclusivos: os que permaneceram na mesma ocupação nos dois períodos e os que mudaram de ocupação. Denotemos por A_1 e A_2 as distribuições ocupacionais na origem e destino, respectivamente, e por I_1 $(i_1 = 1, 2)$ e I_2 $(i_2 = 1, 2)$ as categorias em cada distribuição.

Se, em seguida, contarmos quantas observações em cada grupo correspondem a cada categoria, poderemos produzir uma seqüência de freqüências observadas $(x_{i_1 i_2})$ analogamente classificadas em uma das quatro células de uma tabela bidimensional (2x2), que, por sua vez, pode ser transformada em uma tabela de probabilidade $p_{i_1 i_2}$ $(i_1 = 1, 2;$ $i_2 = 1, 2)$, com

$$\sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} = 1$$

e

$$p_{i_1 i_2} = m_{i_1 i_2}/N$$

onde

$$m_{i_1 i_2} = E(x_{i_1 i_2})$$

A condição de independência dos eventos A_1 e A_2 implicaria que $p_{i_1 i_2} = (p_{i_1 .})\,(p_{. i_2})$, onde o ponto indica a soma de probabilidades na linha (coluna) cujo índice foi suprimido. Logo, dada a identidade acima e supondo

(13) Consulte-se, entretanto, o influente artigo de Ben Porath (1970), além dos trabalhos mais recentes de Heckman (1976), Rosen (1976) e Haley (1976), entre outros. Para uma crítica mordaz e brilhante das suposições do modelo de maximização de utilidade na teoria neoclássica, ver Hollis e Nell (1975). Ainda assim, para aqueles que desejam, será sempre possível interpretar nossos resultados — heuristicamente no contexto da teoria da maximização de utilidade.

(14) Como Barr e Hall (1973, p. 14) salientaram, "a discussão de McFadden e outros... pode provocar algumas confusões neste ponto". Usuários potenciais devem-se precaver contra programas de computadores listados na nota de rodapé (4) de MacFadden (1976a). Eles só serão operacionais na presença de informações sobre os atributos de cada escolha.

(15) Esta seção apóia-se nos capítulos 2 e 6 de Fienberg (1977) e no capítulo 2 de Bishop, Fienberg e Holland (1975).

independência entre as duas variáveis, teríamos a seguinte expressão para a freqüência esperada em cada célula da matriz:

$$m_{i_1 i_2} = N(p_{i_1 .})(p_{. i_2}) \qquad (1)$$

Substituindo na equação acima a proporção observada na amostra, x_{i_1}/N, como uma estimativa de p_{i_1} e a proporção correspondente, x_{i_2}/N, para $p_{. i_2}$, obtemos a expressão usual do valor esperado de probabilidade na célula $(i_1 \; i_2)$, no modelo de independência:

$$\hat{p}_{i_1 i_2} = \frac{x_{i_1} \; x_{. i_2}}{N} \qquad (2)$$

Transformando em logaritmos ambos os termos da equação, teríamos:

$$\log \hat{p}_{i_1 \; i_2} = \log x_{i_1} + \log x_{. \; i_2} - \log N \qquad (3)$$

Pensando em termos de uma tabela I_1 x I_2, com I_1 linhas e I_2 colunas, revela-se na equação (3) uma similaridade bastante grande com a notação convencional em análise de variância. Acompanhando a idéia de que a variação total entre duas variáveis independentes pode ser decomposta na soma de um efeito global, um efeito-linha e um efeito-coluna, o parâmetro $p_{i_1 i_2}$ pode ser expresso da seguinte maneira[16]:

$$\log p_{i_1 i_2} = u + u_1 (i_1) + u_2(i_2) \qquad (4)$$

onde u é a média global dos logaritmos dos valores esperados das probabilidades,

$$u = \frac{1}{I_1 \; I_2} \sum_{i_1 = 1}^{I_1} \sum_{i_2 = 1}^{I_2} \log p_{i_1 \; i_2}$$

$u + u_1(i_1)$ é a média dos logaritmos dos valores esperados nas células I_2 no nível i_1 da primeira variável:

$$u + u_1 (i_1) = \frac{1}{I_2} \sum_{i_2 = 1}^{I_1} \log p_{i_1 \; i_2}$$

e da mesma forma para o nível i_2 da segunda variável:

$$u + u_2 (i_2) = \frac{1}{I_1} \sum_{i_1 =}^{I_1} \log p_{i_1 \; i_2}$$

Como $u_1 (i_1)$ e $u_2 (i_2)$ representam desvios da média global u, tem-se, adicionalmente, a seguinte normalização:

$$\sum_{i_1}^{I_1} u_1(i_1) = \sum_{i_2}^{I_2} u_2(i_2) = 0 \qquad (5)$$

Considerando uma possível interação entre as duas variáveis, deveríamos acrescentar um "termo de interação" ao modelo de independência da equação (4). Neste caso:

$$\log p_{i_1 i_2} = u + u_{1(i_1)} + u_2 (i_2) + u_{12(i_1 i_2)} \qquad (6)$$

onde

$$u_{12(i_1 i_2)} = \log p_{i_1 i_2} - \left[\frac{1}{I_2} \sum_{i_2=1}^{I_2} \log p_{i_1 i_2} + \right.$$

$$\left. + \frac{1}{I_1} \sum_{i_1=1}^{I_1} \log p_{i_1 i_2} + \frac{1}{I_1 I_2} \sum_{i_1=1}^{I_1} \sum_{i_2=1}^{I_2} \log p_{i_1 i_2} \right]$$

- agora, além das restrições em (5), temos:

$$\sum_{i_1=1}^{I_1} u_{12} (i_1 i_2) \qquad \sum_{i_2=1}^{I_2} u_{12} (i_1 i_2) = 0 \qquad (7)$$

O modelo na equação (6) é uma descrição completa da tabela 2 x 2: as probabilidades estimadas para cada célula são iguais às proporções observadas. Visto que o modelo tem tantos parâmetros quanto o número de células na tabela, este é chamado um modelo *saturado* (ou de "full rank").

(16) Ver Fienberg (1977, p. 14).

Em geral, uma amostra de observações sobre q variáveis aleatórias discretas, $A_1 \ldots A_Q$, pode ser organizada em uma tabela de freqüências, $I_1 \times I_2 \ldots \times I_Q$, correspondendo a uma organização similar dos valores esperados nas células $(i_1, i_2 \ldots i_Q)$;

$$m_{i_1 i_2 \ldots i_q} \quad \begin{matrix} i_1 = 1 \ldots I_1 \\ i_2 = 1 \ldots I_2 \\ \vdots \\ i_q = 1 \ldots I_Q \end{matrix}$$

Transformando estes valores em logaritmos e organizando os elementos $\log p_{i_1 \ldots i_q}$ em uma anotação ANOVA de análise da variância, podemos escrever:

$$\begin{aligned} \log p_{i_1 \ldots i_q} &= u + u_1(i_1) + u_2(i_2) + \cdots + u_Q(i_q) \\ &+ u_{12}(i_1 i_2) + \cdots + u_{Q-1,\,Q}(i_{q-1}, i_q) \\ &+ u_{123}(i_1 i_2 i_3) + \cdots + u_{Q-2,\,Q-1,\,Q}(i_{q-2}, i_{q-1}, i_q) \\ &\vdots \\ &+ u_{1,\ldots,Q}(i_1, \ldots, i_q) \end{aligned} \qquad (8)$$

onde todos os termos de $u_1(i_1)$ até $u_{1,\ldots Q}(i_1, \ldots i_q)$ satisfazem as restrições ANOVA[17]:

$$\begin{aligned} &u_1(.) = u_2(.) = \cdots = u_q(.) = 0 \\ &u_{12}(i_1, .) = 0; \; u_{12}(., i_2) = 0; \; \cdots \; u_{Q-1,\,Q}(., q) = 0 \\ &u_{123}(i_1, i_2, .) = 0; \; \cdots \; u_{Q-2,\,Q-1,\,Q}(., q-1, q) = 0 \\ &u_{1,\ldots,Q}(i_1, \ldots i_{q-1}, .) = 0; \; \cdots \; u_{1,\ldots,Q}(., i_2, \ldots, i_q) = 0 \end{aligned} \qquad (9)$$

Aqui, como antes, os parâmetros $u_1(i_1)$ etc. têm a interpretação usual em análise de variância: u indica um efeito geral, $u_1(i_1)$ indica um efeito de interação de segunda ordem entre A_1 e A_2 (nos "níveis" i_1 e i_2, respectivamente), independente de variação nas variáveis restantes; $(\ldots)$ $u_{1,\ldots,Q}(i_1, \ldots i_q)$ indica uma interação de ordem q entre $A_1 \ldots, A_Q$ nos "níveis" $i_1, \ldots I_q$, respectivamente, etc.

Como observou Fienberg, o que torna esta abordagem particularmente interessante é o fato de que "cada termo subscrito u no modelo loglinear geral pode ser expresso como uma combinação linear dos logaritmos dos valores esperados (ou, de forma equivalente, dos logaritmos das probabilidades), onde os pesos ou coeficientes empregados na combinação linear somam zero"[18].

Por outro lado, sendo as observações provenientes de distribuições multinominais independentes, pode-se também demonstrar a existência e, na maioria dos casos, a unidade do estimador de máxima verossimilhança (*EMV*), $\hat{m}$, do valor esperado da freqüência, m, em cada célula de tabela descrita pelo loglinear correspondente (ou, equivalentemente, o *EMV* $\hat{p}$ de p)[19].

Postas em conjunto, estas condições (teoremas) dão-nos estimativas para todos os termos u no modelo geral. Isto, no entanto, não basta. Visto que um determinado resultado do modelo depende de uma combinação simultânea dos termos u, há necessidade também de se conhecer a sua distribuição *conjunta*. Para tanto, empregamos o seguinte fato: em uma grande amostra, caso os valores $\{x_{i_1}, \ldots, x_{i_q}\}$ sigam um modelo

(17) Ver Nerlove e Press (1976, p. 8-9).

(18) Fienberg (1977, p. 70). Tais combinações lineares são conhecidas como contrastes lineares.

(19) Este resultado pode ser estendido a outras distribuições. Ver Fienberg (1977, Apêndice II, p. 131-134) ou Bishop, Fienberg e Holland (1975, capítulo 13, p. 455-56), para uma mais completa apresentação dos experimentos. É importante notar que em todas, exceto nas tabelas 2 x 2, alguns dos EMV precisam ser estimados por procedimentos interativos.

de amostra multidimensional, a distribuição conjunta dos contrastes lineares correspondendo a cada um dos termos *u* no modelo, como definido pela restrição (9), é aproximadamente normal multivariada com média e variância conhecidas (por exemplo, podendo ser estimadas pelas técnicas de *M-V*)[20].

3. Testes de Significância do Modelo Loglinear Geral

A expressão (6) é uma representação completa da tabela de mobilidade simples $I_1 \times I_2$, geralmente aplicada nos estudos de mobilidade; a expressão (8), por outro lado, é a generalização das tabelas multidimensionais.

Como vimos, os termos *u* em um modelo loglinear refletem interações ordenadas entre as variáveis da tabela. As hipóteses referentes à estrutura de uma tabela $I_1 \times I_2 \times \ldots \times I_q$ podem, por conseguinte, ser simplesmente introduzidas alterando-se a especificação do modelo correspondente à tabela em questão. Mais precisamente, é possível impor uma estrutura aos dados pelo simples mecanismo de anular no modelo geral alguns dos efeitos de interação, (isto é, os termos *u* de segunda ou maior ordem) ou todos eles. De fato, isto pode ser obtido modificando-se as restrições ANOVA indicadas em (9). As estimativas de máxima verossimilhança das probabilidades do novo modelo podem, então, ser computadas, pois dessa forma pode-se demonstrar que, na maioria dos casos, as novas restrições não afetam a concavidade global da função de verossimilhança conjunta[21].

Existem procedimentos baseados na distribuição central qui-quadrada para testar, em grandes amostras, a adequação de qualquer modelo hipotético *versus* a alternativa de modelo saturado (cujas estimativas como observamos há pouco, são exatamente iguais aos valores observados). Define-se a razão de verossimilhança como:

$$\lambda = V_R/V_S$$

onde V_R é o valor máximo da função de verossimilhança no modelo restrito e V_S o valor correspondente no modelo irrestrito. Assim, caso a hipótese nula seja correta e o tamanho global da amostra importante[22], a quantidade $\lambda^* = -2 \log \lambda$ possuirá uma distribuição qui-quadrada aproximada, com graus de liberdade dados por[23]:

(20) Ver o teorema 5.1 em Fienberg (1977, p. 71-72) ou o teorema 14. 6-2 em Bishop, Fienberg e Holland (1975, p. 493-494). Este resultado é válido também para algumas outras distribuições.

(21) Isto só é inteiramente válido nos casos em que não há células vazias na tabela, devido, não à estrutura dos dados, mas à natureza do processo de amostragem. Ver teorema 7 em Nerlove e Press (1976, p. 38-42).

(22) Não existe uma definição clara para a expressão "suficientemente grande". Em regra geral, a maioria das aplicações interpreta-a considerando o tamanho da amostra como sendo de pelo menos dez vezes o número de células na tabela (Fienberg, 1977, p. 37), ou que a menor freqüência prevista em qualquer célula avizinha-se de 5 (Blalock, 1972, p. 285; também, Reynolds, 1977, p. 159). Geralmente, quanto menor o número de células e quanto mais próximos forem os totais marginais em cada linha, menor poderá ser o número total de observações **N**. Sempre que houver margens de dúvidas quanto à adequação da aproximação, é aconselhável fazer correções visando à continuidade. Entretanto, como Blalock salientou: "Correções visando à continuidade não podem ser facilmente realizadas no caso de tabelas gerais de contingência. Caso o número de células seja relativamente grande e caso somente uma ou duas células tenham freqüências previstas de 5 ou menos, então, aconselha-se, geralmente, dar seqüência aos testes qui-quadrados sem se preocupar com tais correções" (Blalock, 1972, p. 286).

(23) Alternativamente, podemos empregar a estatística qui-quadrada de Pearson assim obtida:

$$X^2 = \sum_{i=1}^{Q} \frac{(\text{estimativas irrestritas} - \text{estimativas restritas})^2}{\text{Estimativas irrestritas}}$$

onde **Q** é o número de parâmetros no modelo. X^2 é assintoticamente equivalente a

g. 1. = (número de restrições no modelo saturado) — (número de restrições no modelo restrito).

No modelo geral da equação (8), sujeita à restrição (9), os termos *u* para interações hierarquicamente superiores (no limite, o de enésima ordem mostrado na última linha da expressão (8)) representam desvios em relação aos termos de ordem hierárquica inferior. Conseqüentemente, ao limitarmos nossas hipóteses a modelos nos quais os termos de ordem superior venham sempre acompanhados dos termos correspondentes às ordens inferiores, estaremos formando um conjunto hierárquico de hipóteses(24). Para tal conjunto, as estatísticas da razão de verossimilhança são independentes umas das outras, a cada "nível" de hierarquia. É possível, portanto, adotar o mesmo procedimento para testar, seqüencialmente, a importância de uma sucessão de hipóteses, e isto sem comprometer o poder do teste(25). Adicionalmente, e de maneira mais geral, pode-se dizer que, se entre dois modelos lineares o modelo II contém somente um subconjunto dos termos contidos no modelo I (ou seja, se II estiver inserido em I), a estatística λ^* do modelo II pode ser dividida em duas partes complementares. O total pode ser decomposto de forma análoga à soma dos quadrados dos resíduos no modelo usual de análise de variância: (a) um cálculo da distância das estimações $\{\hat{m}_i\ (II)\}$ daquelas obtidas no modelo I, $\{\hat{m}_i(I)\}$; (b) um cálculo da distância entre as estimações do modelo I e os valores efetivamente observados nas células da tabela $\{\overline{x_i}\}$. Logo:

$$y^*(II) = -2\{y(\hat{m}(II)) - V(\bar{x})\}$$

$$= 2\{V(\hat{m}(II)) - V(\hat{m}(I))\} -$$

$$- 2\{V(\hat{m}(I)) - V(\hat{x})\} =$$

$$= \lambda^*\{(II)/(I)\} + \lambda^*(I)$$

onde $\lambda^*\{(II/(I)\}$ é o valor de λ^* no modelo II, *condicional* nos resultados do modelo I. Pode-se demonstrar que, sendo λ^* (II) e λ^* (I) assintoticamente distribuídos como qui-quadrado com *V(ii)* e *v(i)* graus de liberdade, respectivamente, então $\lambda^*\{(II)/(I)\}$ é também assintoticamente qui-quadrado com *v(ii)* — *v* (*i*) graus de liberdade(26). Caso, por exemplo, os dois modelos difiram somente por um único termo *u*, um teste de diferença nas estatísticas de igualdade entre os dois modelos pode ser feito estimando-se um modelo com e sem este termo *u* e comparando o valor de λ^* $D = \lambda^*(II) - \lambda^*(I)$ com a tabela de valor da distribuição de qui-quadrado apropriado(27).

λ,entretanto, ao fazermos os testes, desde a forma mais geral do modelo até a mais restrita, é em geral verdade que λ (modelo mais restrito) $>$ λ (modelo menos restrito), resultado este que não se aplica a X^2. Ver Fienberg (1977, p. 49) e também Reynolds (1977, capítulo 6).

(24) Assim, em um conjunto hierárquico, u_{123} não pode ser incluído no modelo a menos que u_{12} e u_{13} estejam todos no modelo. Ver Bishop, Fienberg e Holland (1975, p. 67-68).

(25) O trabalho seminal sobre métodos hierárquicos de testagem em tabelas de contingência multidimensionais é o estudo de Goodman (1970). Davis (1974) tem uma apresentação mais acessível.

(26) Ver teorema 149.8 no livro de Bishop, Fienberg e Holland (1975, p. 525). Este resulsultado é análogo ao teste-F para regressões múltiplas e possui uma interpretação similar.

(27) Um teste alternativo proposto por Wald pode ser calculado a partir dos valores estimados pelo modelo **restrito**, unicamente. Ao contrário, o cálculo da razão de verossimilhança requer estimativas dos modelos restrito e total. Por este motivo, o teste de Wald foi empregado em várias aplicações; por exemplo, em Growder e Grob (1975). Entretanto, como foi demonstrado por Hauck e Donner (1977), o teste de Wald possui as seguintes características consideradas inconvenientes: "(1) para qualquer tamanho de amostra, as estatís-

...

Sendo que a distribuição conjunta do *EMV* dos termos u num modelo loglinear é assintoticamente normal multivariado, um teste da hipótese $H_o : \hat{u}_k = u^o{}_k$ pode empregar o fato de que a estatística

$$\frac{\hat{u}_k - u^o{}_k}{(\text{Var}\ (\hat{u}_k))^{1/2}}$$

é distribuída assintoticamente como a distribuição "t" de Student[28].

4. A Transformação Logística

Apesar do alcance e da facilidade de manipulação, o modelo loglinear indicado nas equações (8) e (9) não satisfaz nossos objetivos. Interessa-nos em particular a ocorrência de mudanças relativas na distribuição dos dois grupos ocupacionais entre os anos terminais de um determinado período; ou seja, a mudança na distribuição A_2 com relação a A_1.

Nas análises de mobilidade, a partir de uma distribuição inicial, digamos de A_1, são conhecidos os valores das freqüências marginais em cada linha da tabela (m_1 e $m._2$); o que nos interessa é a freqüência relativa de casos em cada uma das colunas. Dito em outras palavras, o número de trabalhadores em cada ocupação de origem é conhecido — o que varia é a sua distribuição no ano final do período. Observe-se, no entanto, que, neste caso, sendo fixo o número de observações em cada linha, devemos nos referir a um outro modelo amostral. Mais especificamente, considera-se uma amostra de N_1 indivíduos selecionados da primeira categoria da variável A_1, de N_2 indivíduos selecionados da segunda categoria etc. Pode-se, então, formar uma tabela contando-se quantos dos indivíduos selecionados pertencem, também, à primeira categoria da variável A_2, à segunda categoria etc. Neste caso, as observações paritárias $\{A_1 = i_1, A_2 = i_2\}$ na i_1-ésima categoria de A_1 e na i_2-ésima categoria de A_2 são variáveis multinominais independentes, assumindo valores $i_1 = 1,.\quad I_1$; $i_2 = 1\quad .,I_2$ sendo

$$P_r\ (A_1 = i_1,\ A_2 = i_2) \quad = P_{i\ (i\ )}{}_{2\ 1}, \text{ para}$$

$n_i = 1,2,..., N_i$ e $\sum_{i_2} \quad P_{i\ (i\ )}{}_{2\ 1} = 1$[29].

Evidentemente existe na variável A_1 tantas probabilidades deste tipo quanto categorias.

O "logit" da i-ésima linha de uma tabela assim definida — supondo, para simplificar,

---

ticas do teste de Wald tendem a zero na medida que a distância entre valor estimado do parâmetro e o valor nulo aumentam; (2) o poder do teste de Wald diminui ao nível de significância para alternativas (do valor dos parâmetros) muito distantes do valor nulo (p. 851). Existe, portanto, uma tendência a aceitar a hipótese nula em situações em que o valor da razão de verossimilhança indicaria rejeição.

(28) Ver teorema 14.3-5 no livro de Bishop, Fienberg e Holland (1975, p. 471). Um resultado idêntico pode ser obtido a partir do modelo logístico. Para isto, vez McFadden (1972, Lemma 6, p. 22). O teste é de certa forma similar ao teste-t da teoria de regressão múltipla. Observe, entretanto, que, no modelo loglinear geral, a relação entre duas variáveis é expressa não por meio de um parâmetro único mas por meio de um vetor de parâmetros: $u_1\ (\ ), u_{12}\ (\ ), u_{123}\ (\ ),\ldots, u_1, \ldots\ Q\ (\ )$.

(29) Empregamos letras maiúsculas para indicar que estas probabilidades diferem das correspondentes ao modelo linear geral anteriormente apresentado. Para as últimas, $\sum_{i_1\ i_2} P_{i\ i} = 1$. Estes esquemas de amostragem são equivalentes ao que Manski e Lerman (1977) chamam de amostragem exógena. "Em uma amostragem exógena, o analista define um tomador de decisão t caracterizado pelos atributos Z_t de acordo com a densidade de g e depois observa a escolha feita por este tomador de decisão entre as alternativas do conjunto de escolhas C" (p. 1979).

que cada variável tenha apenas duas categorias — é dado por:

$$L_{i_1} = \log \frac{P_{1(i_1)}}{1 - P_{1(i_1)}} = \log \frac{m_{i_1 1}}{m_{i_1 2}}$$

Observe-se que, sendo N_i fixo, o lado direito da equação acima pode ser reescrito como:

$$L_{i_1} = \log m_{i_1 1} - \log m_{i_1 2} = \log p_{i_1 1} - \log p_{i_1 2}$$

Portanto, dada a transformação loglinear das probabilidades de uma tabela de quatro células (expressa pela equação (6)), o "logit" para a *i*-ésima linha pode ser expresso como:

$$L_{i_1} = u_{2(1)} - u_{2(2)} + u_{12(i_1)} - u_{12(i_1 2)} = 2u_{2(1)} + 2u_{12(i_1)}$$

uma vez que, pelas condições (5) e (7), $u_{2(1)} = -u_{2(2)}$ e $u_{12(i_1)} = -u_{12(i_1 2)}$. Fazendo-se com que $\alpha = 2u_{2(1)}$ e $\beta = 2u_{12(i_1)}$, o "logit" pode, finalmente, ser expresso como:

$$L_i = \alpha + \beta \qquad (10)$$

assim

$$P_{1(1)} = P_r \{A_2 = 1 \mid A_1 = 1\} = \frac{1}{1+\exp\{-(\alpha+\beta)\}} \qquad (11)$$

que se encontra em uma função de freqüências acumulada na forma logística em $(\alpha+\beta)$, para uma variável categórica dicotômica[30].

Na verdade, já que estamos interessados na ocorrência de mobilidade no *interior de uma determinada linha* da tabela (por exemplo, a partir de uma determinada ocupação de origem), consideramos correto transformar as variáveis ocupacionais (A_1 e A_2) em uma única variável dicotômica de mobilidade. Ou seja, *para uma determinada categoria* i_1 *de* A_1, definimos uma nova variável Y_1 (i=1,2,...,I_1), sendo igual a 0 (zero) quando o trabalhador se encontra no mesmo grupo ocupacional em A_2 e em A_1 e igual a 1 em caso contrário. Defina-se $\{y_i\}$ de forma que $\{y_i\} = (y_1, y_2, \ldots, y_N)$ represente o conjunto de observações na amostra N_1. O conjunto $\{y_i\}$ é composto de variáveis binares aleatórias e independentes, podendo assumir valores 0 ou 1, sendo que:

$$P_r(y_i = 1) = P_i, \; i = 1, 2, \ldots, N_i \qquad (12)$$

onde os valores de P_i são distribuídos como na equação (11).

É provável, entretanto, que esta probabilidade seja influenciada por outros fatores, além das interações entre a ocupação anterior e atual. Suponha-se que queiramos considerar a relação existente entre y_i e as demais variáveis A_3, A_4,. A_q, que podemos imaginar como sendo um vetor de atributos pessoais e de características dos trabalhos da população representada em A_1. Já que y_i refere-se unicamente à categoria i de A_1, podemos ordenar as demais variáveis A em um novo conjunto de variáveis, cujos valores dependerão do fato de o trabalhador ter pertencido ao grupo ocupacional i em A_1. Para o enésimo indivíduo da amostra, chamamos este *vetor* de Z_i, (i = 1,2,... I_1), onde, para cada i, $Z_i = Z_{i1}, \ldots, Z_{ik}$ correspondendo aos valores reordenados do vetor da variável A_Q[31].

(30) A demonstração acima foi, em parte, tirada do livro de Nerlove e Press (1973, p. 33-35). Ver também Bock (1975, p. 528-30) e Cox (1970, p. 19-23). Observe-se que os ... parâmetros α e β deveriam ser divididos em partes iguais para que se transformassem novamente no formato-**u** do modelo loglinear.

(31) Observe-se que se tivéssemos que formar uma tabela 2 x I_1 x...x I_k correspondente

Indicando uma combinação particular de $y_i - Z_{i1} - \ldots - Z_{ik}$ como a célula $i_y,\ldots,i_k$ na tabela de contingência multidimensional $I_y \times \ldots \times I_k$, *e, supondo que as variáveis Z no vetor Z são condicionalmente independentes dado* y_i, podemos definir esta tabela por meio do seguinte modelo loglinear:

$$\begin{aligned}\log P_{i_y \ldots i_k} &= u + u_{Y(i_Y)} + \\ &+ u_{1(i_1)} + u_{2(i_2)} + \ldots \\ &+ u_{K(i_K)} + u_{Y1(i_Y, i_2)} + \\ &+ u_{Y2(i_Y, i_2)} + \ldots + \\ &+ u_{YK(y,\ k)} \qquad (13)\end{aligned}$$

sujeito às restrições,

$$u_{Y(.)} = u_{1(.)} = u_{K(.)} = 0$$

$$u_{Y(i_y,\ )} = 0;\ u_{Y1(.,i_1)} = 0;\ldots;$$

$$u_{YK(.,i_K)} = 0$$

neste caso,

$$\begin{aligned}L_{i_1 \ldots i_k} &= \log \frac{m_{i_y = 1,\ i_2 \ldots i_k}}{m_{i_y = 2,\ i_1 \ldots i_k}} = \{u_{Y(1)} - u_{Y(2)}\} + \\ &+ \{u_{Y1(1,i_1)} - u_{Y1(2,i_1)}\} + \{u_{Y2(1,i_2)} - u_{Y2(2,i_2)}\} + \\ &\ldots + \{u_{YK(1,i_k)} - u_{YK(2,i_k)}\} \\ &= 2\{u_{Y1(1,i_1)} + u_{Y2(1,i_2)} \ldots + u_{YK(1,i_k)} \qquad (14)\end{aligned}$$

Uma vez mais, à esquerda da equação (14) o símbolo indicando a variável *Y* (mobilidade) é suprimido já que aí se representa simplesmente a razão entre os logaritmos de probabilidade de mudança (*LPM*). Como foi definido, a soma, na última linha da equação (14), de todos os valores de Y_i é zero.

O modelo (14) é logístico e determina a existência de efeitos aditivos na razão dos *LPM* devido à ação simultânea de qualquer par ou conjunto de variáveis *Z*. Em outras palavras, o modelo determina uma relação linear entre logaritmo das chances da variável dependente dicotômica Y_i e o vetor Z_i de observações na amostra N_1. O conjunto das variáveis independentes. Agora, suponhamos que:

$$U_{Y1}(1, i_1) = \alpha_1 = 2\{Z_i(1)'\ \Upsilon(1)\}$$
$$U_{Y2}(1, i_2) = \alpha_2 = 2\{Z_i(2)'\ \Upsilon(2)\}$$

$$U_{YK}(1, i_k) = \alpha_k = 2\{Z_i(k)'\ \Upsilon(k)\}$$

onde $Z_i(1)$, $Z_i(2),\ldots,Z_i(k)$ são cada um dos vetores de *k* variáveis independentes, e $\Upsilon(1)$, $\Upsilon(2),\ldots,\Upsilon(k)$ são vetores de pesos a serem avaliados. Assim, a partir da equação (14), e em analogia com a (10), obteríamos:

$$\begin{aligned}L_{i_1 \ldots i_k} &= \alpha_1 + \alpha_2 + \ldots + \alpha_k \\ &= 2\{Z_i(1)'\ \Upsilon(1) + \ldots + \\ &+ Z_i(k)'\ \Upsilon(k)\} = Z^{*\prime}{}_i\ \Upsilon^* \qquad (15)\end{aligned}$$

onde

$$Z^{*\prime}{}_i = 2\{Z_i(1)'\ \Upsilon(1),\ldots,Z_i(k)'\ \Upsilon(k)$$

e $\Upsilon^* = 2\{\Upsilon(1).\ .,\ \Upsilon(k)\}$

$$\begin{aligned}P_{i_1 \ldots i_k}(1) &= P\{A_2 = 1 \mid A_1 = 1, Z_1 = \\ &= i_1,\ldots,Z_k = i_k\} = \qquad (16)\end{aligned}$$

$$= \frac{1}{1 + \exp Z^{*\prime}{}_i\ \Upsilon^*}$$

Note que $\sum_{j=1} Z^{*\prime}{}_{ij}\ \Upsilon^* = 0$

... aos dois valores das variáveis Y_i e aos valores $i_1,\ldots i_k$ das variáveis $Z_1,\ldots Z_k$ os valores marginais das colunas para cada uma das variáveis Z seriam exatamente iguais aos respectivos valores marginais da tabela 2 x I_3 x...x I_3x... xI_Q dos valores i_1, $i_2,\ldots i_q$ das variáveis $A_1, A_3,\ldots,A_q$. Em outras palavras, a tabela anterior seria uma versão dividida (repartida) da última.

O modelo indicado na equação (14) pode ser facilmente estendido a situações em que, para cada camada da tabela, existam mais do que duas categorias da variável dependente. Isto provavelmente ocorre no caso de dados experimentais em que, a partir de uma determinada ocupação no ano de origem, o trabalhador possa (em teoria) passar para qualquer uma das outras alternativas de I_y[32]. Neste caso, o "logit" da i-ésima linha de y_i poderia ser definido como:

$$L_{i_1,\ldots,i_k} \frac{\underset{v}{m_i = 1, i_1,\ldots,i_k}}{\underset{v}{m_i = j, i, \ldots,i_k}} = \frac{P_{i_1,\ldots,i_k}(1)}{\sum_j P_{i_1,\ldots,i_k}(j)} \quad (17)$$

onde $i, j = 1,\ldots, I_y$. A expressão logística para o P_i's, análoga à (16), seria $L_{i_1,\ldots i_k} = Z^*_{ij}\,\Upsilon^*$, sendo que

$$P_{i_1,\ldots,i_k}(1) = \frac{\exp Z^*_{ij}{}'\,\Upsilon^*}{\sum_j \exp Z^*_{ij}{}'\,\Upsilon^*} \quad (18)$$

e

$$\sum_{j=1}^{I_y} Z^*_{ij}{}'\,\Upsilon^* = 0$$

A equivalência entre os modelos (14) e (16) (ou suas contrapartes politômicas) tem uma importante implicação na estimação de máxima verossimilhança nas tabelas de contingência.

Dado que qualquer transformação dos estimadores de *MV* são estimações de *MV* da mesma transformação nos parâmetros, existe a possibilidade de se estimar os efeitos logísticos mediante a resolução da equação (16) — ou, alternativamente, da (18). Isto é até desejável, pois, caso tenhamos de considerar um refinamento cada vez maior dos valores das variáveis categoriais explanatórias, aproximar-nos-emos de uma situação na qual as variáveis explanatórias serão, de fato, tratadas como contínuas. Nesta situação, se o tamanho da amostra for fixo, é possível que as tabelas correspondentes apresentem linhas com freqüências marginais nulas ou quase-nulas. Isto representaria uma clara contradição à hipótese de grandes amostras necessária para derivar as propriedades dos *EMV* dos termos *u* nos modelos loglineares das equações (14) ou (17)[33]. No caso da formulação logística (equação (16) ou (18)), é possível, entretanto, avaliar os *EMV* mesmo em situações nas quais algumas células da tabela geral de contingências sejam nulas ou quase-nulas[34]. À semelhança do modelo loglinear, a teoria de estimação de máxima verossimilhança estabelece que tais estimações são consistentes; isto é, convergem em direção aos valores da população quando o tamanho da amostra torna-se infinitamente grande e a distribuição normal multivariada aproxima-se da distribuição conjunta dos estimadores, com média igual ao va-

(32) Ou seja, o subscrito **i** em Y_i pode ter valores de 1 a I_y.

(33) Neste caso, o teorema citado na nota 20 não teria mais validade. Ver, entretanto, Bishop, Fienberg e Holland (1975, capítulo 5).

(34) Seguindo a apresentação feita em Nerlove e Press (1973), definimos: $P_{n1}{}^* = P_r$ (o trabalhador n^e na amostragem i^e cairá na categoria 1 da variável dependente) onde:

$$P_{n1^*} = \frac{\exp \theta_{n1}}{\sum_j \exp \theta_{nj}} \quad 1, j = 1, \ldots, I_y n = 1, \ldots, N$$

$$\sum_j \theta_{nj} = 0, \text{ e } \theta_{nj} = Z^*{}'_n \; \Upsilon_{j^*}$$

para valores das variáveis $A_3,\ldots A_Q$ Considerando que $v_{ni} = 1$, caso o trabador n^e da amostra i^e esteja na categoria 1 da variável dependente, e $v_{ni} = 0$, em ca-

...

lor da média na população e com matriz de variância/covariância igual ao negativo do inverso da matriz das segundas derivadas da função de verossimilhança. Adicionalmente, pode-se demonstrar que as variâncias obtidas são as menores possíveis em qualquer estimador não-viesado; isto é, os estimadores de máxima verossimilhança são assintoticamente eficientes. Não existe, entretanto, garantia de justeza ou de eficiência em pequenas amostras(35).

5. Um Exemplo

A implementação do modelo apresentado nas seções anteriores requer um tipo singular de informação, até há pouco, raramente disponível. Tratando-se de uma análise longitudinal, é necessário dispor, além dos dados mais comuns referentes a ocupação. características pessoais, familiares e de trabalho reportados ao momento da entrevista, destas mesmas informações para todos os anos da vida laboral do trabalhador. O exemplo mais completo e melhor elaborado deste tipo de pesquisa foi realizado, em 1965. com sexo masculino, em Monterrey, México. Os dados obtidos são singulares: não somente descrevem com exatidão extraordinária a situação de cada indivíduo nas condições vigentes em 1965, mas incluem, também, uma grande quantidade de informações referentes a seu histórico econômico. Mais importante ainda, a base dos dados contém uma descrição detalhada da trajetória de vida dos indivíduos, com informações sobre as mudanças anuais em mais de trinta variáveis e pesquisa sobre a vida familiar, educação, migrações e histórico ocupacional(36).

A nossa análise sobre mudança ocupacional na amostra de Monterrey, baseia-se em cinco matrizes de transição, de, originalmente, dez grupos ocupacionais. Cada matriz descreve o padrão de mobilidade observado em um intervalo de cinco anos; logo, em seu conjunto, a análise engloba os anos de 1940 a 1965. O método estatístico decompõe o conjunto de probabilidade de transição (P_{ni}^*) que une uma origem comum a distintos destinos ocupacionais, em termos de uma série de vetores (Z_{ni}) de variáveis individuais ou de características do emprego. Como foi discutido nas seções anteriores, consideramos que, para cada ocupação, a relação entre Z_n^i e P_{ni}' é descrita por uma função logística de probabilidade do tipo:

$$P_{ni}^* = \frac{\exp \theta_{nj}}{\sum \exp \theta_{nj}}$$

onde Θ é o vetor de parâmetros. Neste caso, θ toma a forma de $\Theta_0 + \Theta_i + \Theta_i + \Theta_k + \sum_{e=1}^{6} \Theta_e Z_e$ (onde $\Sigma_i \theta_i = 0$, e $\Sigma_k \theta_k = 0$). Os símbolos são definidos

---

.. so contrário, a função de log-verossimilhança é:

$$L = \prod_{n=1}^{N} \prod_{i=1}^{I_y} v_{ni}^{P_{ni}^*}$$

com

$$\sum_{i=1}^{I_y} P_{ni}' = 1 \text{ e } \sum_{i=1}^{I_y} v_{ni}$$

A função assim definida é estritamente côncava, com uma matriz hessiana independente das respostas observadas. Conseqüentemente, é possível desenhar programas eficientes para a maximização da função mediante subrutinas iterativas.

(35) Os detalhes do método de máxima verossimilhança aplicados ao "logit" multinominal, incluindo suas propriedades estatísticas, são discutidos em Nerlove e Press (1973), (1976); McFadden (1973); Bock (1975, capítulo 8); entre outros. Com respeito à questão do tamanho mínimo da amostra, não há, em geral, problema algum, contanto que se empregue técnicas MV (ver nota 21).

(36) Esta base de dados foi extensivamente analisada por Bálan, Browning e Jelin (1973) em seu excelente estudo sobre migração, realização profissional e estratificação social em Monterrey. Em uma perspectiva diferente, foi reexaminado por Vieira da Cunha (1980).

como: Θ_0 é uma "média" de Θ quando nenhum outro efeito está presente; θ_i, efeito da experiência geral no nível *i*; θ_j, efeito da experiência profissional no nível *j*; θ_k, efeito da educação no nível *k*; Z_1 é um índice de especialização (para trabalhadores exercendo as profissões manuais na base da pirâmide ocupacional); Z_2 é o índice do histórico sócio-econômico e Z_3 o índice do histórico geográfico; Z_4 e Z_5 medem o tamanho da empresa e Z_6 o setor de emprego[37].

6. O Desempenho Global do Modelo de Mobilidade

Iniciamos nossa discussão sobre os resultados empíricos, examinando o desempenho global do modelo logístico de mobilidade implementado com os dados obtidos no estudo de Monterrey.

Infelizmente, para exame deste problema não podemos empregar, aqui, o método usual em análises de regressão, que, como é sabido, consiste no cálculo do quociente da variação "explicada" pelo modelo para o total da amostra. A "variável dependente" (trajetórias de mobilidade) não é uma quantidade e não possui uma medida exata de variação. É esta falta de uma medida de variação que torna difícil saber até que ponto os resultados empíricos estão ou não em conformidade com as hipóteses teóricas. Um único resultado significante no teste de qui-quadrado indica que as freqüências previstas não estão em conformidade com os dados observados. A rejeição da hipótese nula não deve ser, isoladamente interpretada como uma medida de adequação teórica da classificação adotada (ou seja, da escolha das variáveis). Por este motivo, preferimos não empregar transformações do tipo R^2 para a estatística Υ^* – razão de verossimilhança —: apesar de uma similaridade aparente, os resultados não seriam comparáveis[38].

Decidimos seguir uma orientação diferente. Uma forma simples de testar a adequação do modelo aos dados é medir seu poder de classificação, ou seja, a capacidade do modelo de prever a ocupação de destino de um indivíduo tendo por base, unicamente, os valores de seus atributos prévios, vale dizer, os valores das variáveis "independentes" (predeterminadas) incluídas na análise. O quociente entre o total de mudanças previstas pelo modelo e o total efetivamente observado pode ser interpretado como uma medida do grau de ajustamento do modelo

TABELA 1

PERCENTAGENS DE CLASSIFICAÇÃO[a]
1940 — 1965

Período	Ocupações de Origem			
	5-4-3	8-7-6	9	10
1940-45	77%[b]	—	73%	80%
1945-50	77%	—	71%	83%
1950-55	88%	89%	57%	86%
1955-60	82%	87%	74%	84%
1960-65	94%	90%	75%	90%

OBS.:(a) O grupo 10 corresponde às ocupações manuais hierarquicamente inferiores; o grupo 9, às ocupações manuais semiqualificadas; o grupo 8-7-6, às ocupações industriais; o grupo 5-4-3, às ocupações não manuais de nível inferior e as ocupações 2-1, às ocupações administrativas e gerenciais. Ver Vieira da Cunha (1980, Apêndice I, p. 323-367), para uma descrição mais detalhada das ocupações e uma exposição dos métodos de agregação empregados para reduzir as 98 classificações ocupacionais originalmente reportadas na amostra em, sucessivamente, 39, 10 e 5 grupos ocupacionais.

(b) Percentagem de casos classificados corretamente pelo modelo (avaliado por meio das variáveis independentes).

(37) Não é nossa intenção neste trabalho, explorar os argumentos teóricos que nos levam a tal especificação do modelo. Sobre isso, ver Vieira da Cunha (1980, capítulo 1).

(38) McFadden (1972, p. 23) propõe tal análogo. Theil (1972) e Kohn (1976) proferem medidas de incerteza baseadas no conceito de entropia; elas também estão restritas ao intervalo 0-1.

TABELA 2

QUOCIENTE ENTRE FREQÜÊNCIAS ESTIMADAS(a) E OBSERVADAS, POR OCUPAÇÃO DE ORIGEM

1940-1960

Ocupação/Período		Ocupação				
		10	9	8-7-6	5-4-3	2-1
Oc. 10	1940-45	1.15	0.92	0.64		0.53
	1945-50	1.13	1.03	0.65		0.59
	1950-55	1.10	0.87	0.82		0.42
	1955-60	1.11	1.04	0.49		0.78
	1960-65	1.07	0.98	0.45		0.79
Oc. 9	1940-45	0.70	1.29	0.61		0.97
	1945-50	0.42	1.25	0.69		0.85
	1950-55	0.55	1.23	0.54		0.52
	1955-60	0.64	1.22	0.76		0.60
	1960-65	0.77	1.19	0.43		0.79
Oc. 8-7-6	1950-55	0.79	0.02	1.06		0.80
	1955-60	0.14	0.11	1.08		0.72
	1960-65	0.17	0.67	1.06		0.77
Oc. 5-4-3	1940-45		0.43		1.18	0.73
	1945-50		0.18		1.14	0.72
	1950-55		0.71	0.76	1.08	0.64
	1955-60	0.65		0.23	1.12	0.61
	1960-65	0.68		5-4-3	1.10	0.39

OBS.: (a) Estimados nos valores médios de todas as variáveis independentes.

aos dados[39]. A tabela 1 apresenta esta medida.

Vê-se na tabela que, na maioria dos casos, o modelo prevê corretamente mais que três quartos das mudanças — um resultado bastante satisfatório tendo em vista o contraste entre a estrutura relativamente simples do modelo e a complexidade real do processo de mobilidade ocupacional. Via de regra, as taxas de classificação apresentam uma melhora nos períodos mais recentes; ainda assim, a variação registrada ao longo das décadas dos 40 e 50 é grande e permite-nos questionar a validez de uma correlação entre data da informação e qualidade de resultados. Vale a pena notar que, para todos os períodos, o modelo mostra-se menos eficiente em prever mudanças originais nos trabalhos manuais semiqualifcados incluídos na ocupação 9: para os anos de 1950 a 1955, somente 57% dos casos foram classificados de forma correta. Isto deve-se, em parte, à grande proporção de trabalhadores que permaneceram estáveis naquela exata ocupação e período (67% dos trabalhadores incluídos na ocupação 9 em 1950 estavam incluídos no mesmo grupo ocupacional em 1955) e também à extremamente pequena capacidade de previsão do modelo com relação ao grupo (a proporção estimada de trabalhadores estáveis é de 82%)[40].

De fato, a proporção de trabalhadores estáveis é sistematicamente superestimada em *todos* os períodos e em *todas* as ocupações.

Este resultado pode ser facilmente constatado na tabela 2, que apresenta a percentagem das proporções previstas e observadas para cada uma das mudanças possíveis,

(39) Ver Press e Wilson (1978), Crowder e Grob (1975). Uma terceira abordagem mais geral seria examinar os resíduos da equação, como sugerido em Whitney e Boots (1978, p. 163-164).

(40) Um erro comparável (uma diferença de 15% entre as distribuições previstas e observadas) ocorreu no período de 1940-1945, mas, neste caso, a proporção observada de trabalhadores fixos é de 59%.

dada a ocupação de origem. A superestimação dos elementos da diagonal principal da matriz de transição indica que, caso o processo de mobilidade *fosse reduzido a sua versão modelada*, o número de trabalhadores estáveis superaria o realmente observado. Entretanto, as relações que influenciam qualquer mudança ocupacional são, sem exceção, mais complexas do que as incluídas no modelo.

Se procurássemos abordar o tema da mobilidade a partir da lógica da ação individual e não por seus determinantes estruturais, proavelmente concluiríamos que, sobretudo, as mudanças ocupacionais estão relacionadas com características inerentes aos indivíduos — pessoais e/ou familiares —, mas, destas, algumas não teriam sido correta ou cabalmente delineadas no modelo. Estar "no lugar certo no momento certo", conhecer alguém de influência que preste seu apoio etc. seriam forças não mensuradas mas decisivas no confronto ocupacional — e que pesariam mais do que os elementos da demanda da posição profissional ou capacidade de trabalho.

Sem dúvida, existe alguma verdade nesta afirmação; principalmente se considerarmos tais elementos fortuitos como complementos do processo especificado pelo modelo. Desta forma, eles poderiam influenciar — mas, neste caso, desordenadamente — a trajetória de mobilidade; seu efeito sistemático seria capturado pelos parâmetros das variáveis explicitamente incluídas na análise. Se aceitamos esta visão das limitações do modelo, os achados apresentados na tabela 2 têm uma explicação coerente. Sem a influência adicional dos elementos idiossincráticos, a mobilidade potencial seria menor do que a observada, o que é válido tanto para as mudanças em ocupações de nível inferiores como superiores. Algumas das mudanças observadas refletem circunstâncias particulares que escapam à norma mas não a invalidam[41].

Existem, no entanto, resultados contraditórios na tabela 2. Em certos casos, a proporção observada é duas vezes maior do que a prevista. Isto ocorre, por exemplo, com as mudanças que têm origem nos empregos do escalão ocupacional inferior (representadas pela ocupação 10) e que terminam nas ocupações industriais (8-7-6), nos períodos 1955-60 e 1960-65. O mesmo acontece com as mudanças dirigidas às ocupações não-manuais e administrativas (5-1, em nossa classificação), nos períodos 1945-50, 1950-55 e 1955-60. Em outros casos, a percentagem prevista é inferior a um terço da observada e em alguns exemplos pouco mais de 10%.

Devemos salientar, entretanto, que em todos estes casos as proporções observadas eram, de início, muito pequenas. Em outras palavras, a base de dados que serviu como matéria-prima para as estimações continha poucos casos de trabalhadores que faziam estes tipos de mudança; e alguns poucos casos mal classificados podem, nestas circunstâncias, provocar uma discrepância muito grande entre proporções *previstas* e observadas[42].

Por outro lado, visto que para todos os períodos e em todas as categorias ocupacio-

(41) Devemos salientar que uma tendência mais difícil de ser explicada, na direção ... inversa (isto é, de subestimativa dos alementos na diagonal principal), é o resultado mais comum obtido na aplicação dos modelos de cadeia de Markov a dados de mobilidade intergeneracional. Kogan e McCarthy (1955, p. 60-65) observaram que esta subestimativa aumentava com a idade do trabalhador, e é este o resultado que os levou a propor a diferenciação entre trabalhadores estáveis e móveis. Ver também McFarland (1970) e Sewman (1976) para uma análise destes e outros resultados similares.

(42) Por exemplo, havia somente 5 trabalhadores que durante os anos entre 1950 e 1955 passaram de um trabalho incluído nas ocupações 6, 7 e 8 para uma outra ocupação 9. O modelo estimado levou em conta apenas um destes — deixando de classificar de forma correta quatro de um total de 170, na soma destas ocupações em 1950.

TABELA 3

RAZÕES DE VEROSSIMILHANÇA (λ^*) DOS MODELOS ESTIMADOS

		Número de Observações	$\lambda^{*(a)}$	Graus de Liberdade
Ocupação 10	1940-45	153	161.02	30
	1945-50	199	193.08	30
	1950-55	211	256.72	30
	1955-60	266	332.69	30
	1960-65	292	396.43	30
Ocupação 9	1940-45	97	78.80	30
	1945-50	105	99.60	30
	1950-55	147	192.67	30
	1955-60	201	214.89	30
	1960-65	227	272.91	30
Ocupações 8-7-6	1950-55	170	304.36	27
	1955-60	213	418.06	27
	1960-65	291	522.91	27
Ocupações 5-4-3	1940-45	94	91.19	18
	1945-50	122	142.31	18
	1950-55	177	203.74	18
	1955-60	222	316.53	27
	1960-65	287	546.85	27

OBS.: (a) Em todos os casos, a hipótese nula é rejeitada com probabilidade de erro inferior a 1%.

nais os trabalhadores estáveis formavam o grupo predominante, torna-se óbvio que uma superestimativa da probabilidade de permanência relativamente pequena estaria associada a um número maior de casos classificados de forma incorreta. Quanto maior a proporção de trabalhadores estáveis e a superestimativa desta proporção, maior será a subestimação do número de casos nas demais células na matriz. Tipicamente, é o que ocorre com as mudanças, cuja origem está nas tarefas industriais da ocupação 8-76.

É surpreendente observar que nenhuma destas objeções poderia ter sido detectada por intermédio do escrutínio das estatísticas da razão de verossimilhança (Υ^*) apresentadas na tabela 3. Mesmo permitindo uma única margem de erro de 1%, todos os valores são superiores ao valor correspondente da distribuição qui-quadrada. Em todos os casos, podemos, sem dúvida alguma, rejeitar a hipótese de que o processo de mobilidade em questão poderia ter sido melhor descrito sem referência às variáveis independentes. Como previsto, o valor de Υ^* aumenta com o número de observações mas diminui quando decresce o número de células da tabela.

Em suma, mesmo que a capacidade de previsão do modelo para alguns tipos de mudança pareça bastante pequena, o seu desempenho global mostra-se mais do que adequado.

A proporção de casos classificados de forma incorreta é, de maneira geral, pequena, sendo que em todos os períodos os índices de classificação são razoavelmente bons. O que é ainda mais importante é que os modelos estimados são, sem exceção, estatisticamente significativos. Apesar de a especificação empírica do processo mobilidade ser aleatória e incompleta, os resultados indicam uma estreita relação entre os valores das variáveis independentes e os resultados ocupacionais. Isto não significa que todas as variáveis do modelo tenham a mesma importância qualitativa ou sejam quantitativamente influentes. Para avaliar este impacto isolado, devemos levar em consideração alguns resultados adicionais, o que faremos a seguir.

TABELA 4

RAZÃO DE VEROSSIMILHANCA (λ**) DOS MODELOS ESTIMADOS, COM SUPRESSÃO SISTEMÁTICA DOS EFEITOS DEVIDO ÀS VARIÁVEIS INDEPENDENTES

	Período	Experiência		Experiência Ocupacional		Escolaridade		Especialização		Origem Social Geográfica		Características da Empresa	
		λ**	gl	λ**	gl	λ**	gl	λ**	gl	λ**	gl	λ**	gl
Ocup. 10	1940-45	13.81 (a)	3	13.24 (a)	3	52.82 (a)	3	11.29 (b)	3	10.05 (d)	6	9.92	6
	1945-50	12.64 (b)	3	56.92 (a)	3	35.07 (a)	3	12.82 (a)	3	16.37 (b)	6	8.56	6
	1950-55	5.86	3	13.85 (a)	3	50.63 (a)	3	11.78 (b)	3	14.25 (b)	6	3.26	6
	1955-60	4.13	3	4.67	3	25.44 (a)	3	8.26 (c)	3	18.46 (a)	6	10.25 (d)	6
	1960-65	4.63	3	10.17 (b)	3	37.29 (a)	3	7.95 (c)	3	7.66 (c)	6	15.14 (b)	6
Ocup. 9	1940-45	22.16 (a)	3	23.05 (a)	3	52.67 (a)	3	22.61 (a)	3	2.34	6	14.70 (b)	6
	1945-50	10.36 (b)	3	41.26 (a)	3	67.45 (a)	3	17.55 (a)	3	22.25 (a)	6	6.50	6
	1950-55	47.16 (a)	3	28.62 (a)	3	21.82 (a)	3	48.61 (a)	3	20.71 (a)	6	29.65 (a)	6
	1955-60	13.22 (a)	3	23.11 (a)	3	22.86 (a)	3	30.22 (a)	3	99.24 (a)	6	14.12 (c)	6
	1960-65	34.15 (a)	3	45.61 (a)	3	203.67 (a)	3	29.06 (a)	3	9.09	6	4.51	6
Ocupações 8-7-6	1950-55	7.13 (c)	3	3.95	3	11.56 (b)	3			11.85 (d)	6	3.75	6
	1955-60	19.71 (c)	3	5.36	3	13.52 (a)	3			12.00 (d)	6	12.27 (c)	6
	1960-65	4.08	3	22.68 (a)	3	33.73 (a)	3			16.10 (b)	6	9.14	6
Ocupações 5-4-3	1940-45	6.39 (c)	2	64.86 (a)	2	18.32 (a)	2			10.56 (c)	4	1.72	4
	1945-50	6.65 (c)	2	20.05 (a)	2	64.29 (a)	2			15.06 (a)	4	6.33	4
	1950-55	2.04	2	21.85 (a)	2	não	não			9.00 (d)	4	1.86	4
	1955-60	14.37 (a)	3	76.92 (a)	3	14.04 (a)	3			10.88 (d)	6	7.29	6
	1960-65	15.10 (a)	3	132.88 (a)	3	51.90 (a)	3			25.26 (a)	6	6.38	6

OBS.: 1. A estatística λ** corresponde a duas vezes a diferença nas probabilidades Log entre o modelo completo estimado com os valores dos parâmetros correspondentes substituídos por zeros. Esta estatística é distribuída assintoticamente como qui-quadrada com tantos graus de liberdade quanto coeficientes igualados a zero. Ver NERLOVE & PRESS (1976, p. 46).

2. As letras (a), (b), (c) (d) indicam, respectivamente, valores estatisticamentes significantes ao nível de 1% ou mais, ao nível de 5%, de 10% de 20%.

3. As variáveis não incluídas na análise são representadas por um traço (-) e os casos para os quais a função de probabilidade não poderia ser obtida pela maximização da função de verossimilhança são representados por "não".

7 Efeitos das Variáveis Independentes sobre o Processo de Mobilidade: Significância (Estatística) das Variáveis

Ao avaliar a contribuição individual de cada uma das variáveis sobre o processo de mobilidade, devemos responder a pelo menos duas perguntas: se é possível isolar a influência de uma variável de todas as outras variações contidas nos dados e qual a assovalor e as alterações na probabilidade de um tipo particular de mudança ocupacional. Esta seção responde à primeira destas perguntas.

Tal tarefa não é, de forma alguma, simples. Para começar, devemos reconhecer que o problema de significância estatística envolve uma dupla problemática: sobre a natureza da verdadeira relação entre as variáveis e como avaliá-la, e sobre o que ocorre quando estudamos esta relação por meio de uma amostra da população.

O primeiro aspecto envolve nada menos do que toda a problemática da montagem do modelo, algo que já discutimos anteriormente e que, a esta altura, deve ser considerado como uma premissa. Caso o processo de mobilidade tenha sido inadequadamente especificado, não há esperança de se conseguir bons resultados. Ainda que nossa especificação seja aceitável (e os resultados anteriores parecem indicar que sim), pode ainda ocorrer que, devido às características da amostra, algumas relações não tenham validade ou não possam ser claramente discernidas das outras variações constantes dos dados.

Estas duas últimas problemáticas podem ser testadas graças às nossas hipóteses sobre as distribuições de probabilidade das vaexceto algumas exceções, todas as variáveis definida na seção 3. Um exame dos resultados apresentados na tabela 4 sugere que, salvo algumas exceções, todas as variáveis incluídas na análise estatística estão relacio-

nadas de forma significativa com o processo de mobilidade. Numa comparação global, a formação escolar aparece como a influência mais importante, seguida pela experiência profissional, especialização e experiência geral, em ordem decrescente.

Apesar de as variáveis relativas ao histórico familiar e geográfico e à empresa estarem, na maioria dos casos, relacionadas de forma significativa com a probabilidade de movimento, a força da relação existente entre elas é pequena — principalmente no caso das variáveis das empresas (tamanho do estabelecimento).

Antes de passarmos a uma apresentação mais minuciosa da tabela 4, convém mencionar algumas de suas limitações. Consideraremos duas observações, a nosso ver principais.

Primeiro, deve-se ter em mente que na análise todas as relações são parciais. Qualquer efeito contém, implicitamente, a expressão "todas as outras variáveis são constantes"; na prática, quando se está trabalhando com modelos causais, pode-se querer controlar variáveis diferentes em diferentes etapas da análise. Por exemplo, observamos, nos resultados obtidos anteriormente, que as variáveis do histórico familiar e geográfico não exercem o mesmo tipo de influência sobre a mobilidade exercido pela escolaridade. No entanto, em uma etapa anterior da vida do trabalhador, as origens sócio-econômicas e geográficas influenciaram de forma importante a quantidade (para não se falar da qualidade) de educação recebida[43]. Tanto no caso das variáveis do histórico, como da empresa, este problema apresentar-se-ia sob forma de multicolinearidade entre as variáveis "independentes"[44] Um teste verdadeiro da influência conjunta de todas as variáveis relacionadas implicaria a exclusão, primeiro, das variáveis do histórico e da educação e, em seguida, das variáveis da empresa, de experiência ocupacional e de especialização. É obvio, que, decididamente, rejeitaríamos a hipótese nula em todos os períodos e em todas as ocupações. A questão não está no fato de as variáveis do histórico e da empresa serem irrelevantes. ou, menos importante, na análise do processo de mobilidade, mas, no fato de, em relação às outras variáveis, a sua contribuição adicional parecer menor.

Isto leva-nos a uma segunda limitação: não podemos esquecer o princípio elementar de que os testes de significância são sensíveis ao tamanho da amostra, assim como à força dos efeitos da variável em questão. Ao trabalharmos com amostras extremamente grandes, até o mais insignificante efeito pode ter importância; mas, caso o número estimado de casos na(s) célula(s) referente(s) ao efeito em questão seja pequeno, surgem dúvidas sobre a própria adequação teórica do teste qui-quadrado. Isto significa que deveríamos ser prudentes ao

(43) Felizmente para nós, Balán, Browning e Jelin (1973) trataram extensivamente este assunto no seu estudo sobre os dados de Monterrey. A escolaridade dos pais, assim como a profissão do pai, influenciam de forma significativa a escolaridade do entrevistado, a qual, aliás, também recebe uma influência positiva e significativa da variável tamanho da comunidade de origem. Entretanto, todos os efeitos variam, dependendo da idade do trabalhador. Com ... o tempo, nota-se um declínio nos efeitos da origem sócio-econômica e da comunidade sobre a escolaridade e a proporção da variância da escolaridade "explicada" por estas variáveis passa de 51 para 42% (Balán, Browning e Jelin, 1973, p. 270-274). De acordo com os autores, "o declínio em importância da ocupação do pai é compatível com a hipótese de diminuição em importância dos diferenciais econômicos na determinação das oportunidades educacionais (ibid., p. 274). Isto é devido, principalmente, à expansão do sistema público gratuito de educação. O impacto da origem sócio-econômica na educação é também documentado em uma amostra de aproximadamente 2.500 trabalhadores na cidade do México, por Muñoz, Hernandez e Rodriguez (1978).

(44) Ver Bowles (1972).

fazer comparações entre amostras de tamanhos diferentes. Na verdade, se uma determinada estatística da razão de verossimilhança tiver um valor superior em uma amostra menor — como é o caso, por exemplo, da escolaridade na ocupação 10 para os períodos 1950-55 (N=211, **=50.63) e 1955-60 (N = 266, ** =25.44) — este resultado deve ser vigoroso. Entretanto, pode acontecer que, mesmo em anos próximos a 1965, quando o número total de trabalhadores na amostra aumenta, ainda aí uma célula da matriz de mobilidade possua um número *menor* de elementos que em períodos anteriores. Se esta for a situação, e caso a célula em questão tenha uma importância crítica na determinação da força da relação, os dois resultados não serão realmente comparáveis(45) (46).

Retornando à tabela 4, podemos agora proceder a um cálculo mais detalhado de nossos resultados. Em geral, escolaridade e experiência ocupacional são variáveis mais intensamente associadas aos padrões de mobilidade observados a partir da ocupação 10. A importância deste resultado está na sua associação com as posições na base da pirâmide ocupacional, sendo, portanto, contraditório com a hipótese de "marginalização" veementemente alardeada na década dos 60. Entretanto, para o período 1955-60, a força desta relação diminui substancialmente, a ponto de as variáveis na experiência ocupacional não estarem mais relacionadas de forma significativa com a distribuição de resultados de mobilidade ocupacional. Por outro lado, as características do histórico pessoal, que, em outros períodos, não se relacionam de forma significativa com a mobilidade, ocupam, agora, uma posição de maior destaque. A especialização está, em todos os períodos, ligada de forma significativa ao padrão de mobilidade observado, mas este não parece ser o caso das variáveis de empresa.

Os resultados das mudanças que têm origem na ocupação 9 (segundo escalão da escada ocupacional) não diferem muito entre si, salvo, talvez, no caso do desempenho, geralmente mais influente na variável de experiência ocupacional. O papel especial exercido pelas influências do histórico social e geográfico durante o período 1955-60 pode mais uma vez ser observado; neste caso, tal papel chega a estender-se ao período 1950-55. A experiência geral apresenta no todo uma relação mais estreita com a variável dependente e, para o período 1950-55, supera tanto a variável da educação como a da experiência ocupacional em grau relativo de significância. É importante observar também que, na maioria dos períodos, a força de todas as relações aumenta quando passamos das equações da ocupação 10 para as da ocupação 9.

Isto não é, certamente, o caso dos valores de λ^{**} nas ocupações 8-7-6. No que se refere à mobilidade fora dos empregos industriais, os valores da estatística da razão de verossimilhança para a variável experiência ocupacional, apresentados na tabela 4, não são apenas inferiores mas também insignificantes. Escolaridade, experiência e

(45) Isto é na verdade o que ocorre no exemplo anterior, como se tornará evidente quando voltarmos a nossa atenção ao resultados dos parâmetros individuais e às características t correspondentes.

(46) Poderíamos acrescentar duas outras limitações. A terceira, refere-se ao problema de generalizações a partir de dados amostrais. Neste caso, o teste qui-quadrado presume que os dados provenham de uma amostra multinominal simples e aleatória; mas isso raramente ocorre e, com certeza, não nos nossos dados. Por incluir dados retrospectivos, a amostra possui propriedades desconhecidas em todos os outros anos que não 1965. Adicionalmente, mesmo para este ano, o cálculo amostral baseou-se num complexo sistema de quotas e ponderações que refletem exatamente as características da população, mas distorcem o modelo amostral (sobre isso, ver Balán, Browning e Jelin (1973, Apêndice A). Quarto, os testes de significância devem ser suplementados por estatísticas descritivas que exprimam a magnitude e direção da relação. Isto faremos na seção seguinte.

histórico social e geográfico são, nesta ordem, as variáveis cujas distribuições estão em maior conformidade com os padrões de mobilidade observados[47].

A organização predominante surge, mais uma vez, para mudanças com origem nos empregos administrativos das ocupações 5-4-3. De fato, nestas mudanças e para a maioria dos períodos, a experiência ocupacional é a variável de maior força associativa. As características de escolaridade, histórico social e geográfico e experiência geral são, em ordem decrescente, as mais importantes e sem interrupções para os períodos 1950-55 ou 1955-60.

Em suma, apesar de os resultados observados tenderem a favor das hipóteses postuladas, há provas de heterogeneidade considerável nas relações. As variáveis isoladas para análise estão todas relacionadas de forma significativa com o processo de mobilidade. Entretanto, a força de seu vínculo varia entre as ocupações e durante os períodos de tempo para uma determinada ocupação. Ademais, salvo algumas exceções, não existem tendências nas variações. O vínculo entre escolaridade e mobilidade não diminui visivelmente com o tempo; e nem por isto a conexão entre experiência (geral ou ocupacional) e mobilidade torna-se maior. Tal como se apresentam, as magnitudes e variações nas influências atribuíveis a esta ou aquela variável não são passíveis de ordenação teórica. Para que isto fosse factível seria necessário um esforço adicional que considerasse outros fatores representativos da estrutura econômica — algo que propomos fazer em outro trabalho. Ao considerarmos conjuntamente tanto a escolaridade e especialização como as duas formas de experiência, estas mostram-se muito mais fortemente relacionadas com a mobilidade do que as características do histórico pessoal ou da empresa. Entretanto, visto que o histórico influencia diretamente a educação e que a natureza da empresa determina a qualidade da experiência, estas correlações devem ser um resultado esperado.

Muito mais interessante é o fato de, ao contrário da opinião vigente em alguns círculos, a educação e a especialização não constituírem sempre a única relação mais importante.

8. Efeitos das Variáveis Independentes e Processo de Mobilidade. Significância, Magnitude e Direção dos Parâmetros Individuais

Até o momento, examinamos a força de relações entre distribuições. Estávamos preocupados com a totalidade do modelo de deslocamento ocupacional e não com um resultado em particular.

Não obstante a sua própria importância, esta visão é incompleta, já que uma determinada relação pode apresentar fortes aspectos para uma certa ocupação de destino mas não para outra. Por exemplo, a escolaridade pode estar relacionada de maneira significativa com as mudanças que ocorrem a partir da ocupação 10, mas pode não influenciar a decisão de permanecer na mesma ocupação.

Uma indicação possível, apesar de limitada, da importância de um parâmetro em particular pode ser obtida a partir do resultado exposto a seguir, extraído da teoria de estimação de máxima verossimilhança aplicada a modelos loglineares. Quando o modelo postulado conforma-se com os dados originais, a matriz de variância/covariância dos parâmetros estimados — a partir de amostras grandes — é dada pela negativa do inverso das segundas derivadas do log de V avaliado em $\hat{\beta}$ (o vetor das estimativas do parâmetro), ou seja:

(47) No período 1960-65, a variável experiência ocupacional substitui a experiência geral na lista das variáveis mais fortemente associadas aos padrões de mobilidade. Isto sugere a influência de um possível problema de multicolinearidade entre as variáveis.

$$\sqrt{(\beta)} = - \left[\frac{\partial^2 \text{ Log } V}{(\partial\beta)\ (\partial\beta')} \right]^{-1}_{\beta = \hat{\beta}}$$

Visto que a distribuição de $\hat{\beta}$ é assintoticamente normal as raízes quadradas positivas dos elementos diagonais da matriz (por exemplo, os erros padrões dos coeficientes estimados) podem ser empregados para montar o teste-*t* da teoria padrão de análises lineares de mínimos quadrados.

As limitações do teste-*t* usual são bastante conhecidas. O seu análogo logístico possui as mesmas restrições, se não mais. A hipótese de normalidade é somente assintoticamente válida e os efeitos de multicolinearidade — enviesando para cima as estimativas dos erros padrões — são muito mais acentuados devido ao processo interativo de estimação[48].

Adicionalmente, ao discutirmos os resultados quantitativos, defrontamo-nos com dois problemas suplementares. Visto que a relação entre as variáveis independentes e a probabilidade de mudança não é linear, devemos, primeiramente, especificar um ponto de comparação. Como é de costume fazer, devemos basear nossas comparações em estimações calculadas com as *médias* de todas as variáveis independentes. Em seguida, devemos adotar uma medida da sensibilidade de mudança neste ponto. Mas, a questão de saber qual a medida a ser empregada é mais complexa.

De um lado, algumas variáveis são contínuas mas apresentam, entre períodos, amplitudes de variação diferentes. Isto ocorre com a experiência, experiência ocupacional e escolaridade. Ao contrário, as variáveis restantes apresentam amplitudes idênticas em todos os períodos mas são discretas (dicotômicas). Para as primeiras, podemos definir medidas tais como o efeito marginal ($m = \frac{\partial P_i}{\partial X_j}$) ou a elasticidade de probabilidade (ε), que pode ser interpretada de forma simples e direta[49]. Mas seria errôneo atribuir elasticidade a uma variável descontínua cujos valores são necessariamente 0 ou 1. Para tanto, o que deveria ser avaliado é a diferença nos valores de probabilidade quando uma variável em particular (ou uma combinação de variáveis) estiver ativa ou não.

(48) Ver Kohn (1976), especialmente p. 15. Com extrema multicolinearidade, o inverso de $\sqrt{\beta}$ torna-se numericamente instável e pequenas alterações em P_i de interação para interação causam grandes alterações no valor do inverso.

(49) Sendo $P_i = \frac{\exp\ (Z_i)}{\sum_k \exp\ (Z_k)}$ onde $k = 1 \ldots$ I,...,m é o expoente das possíveis ocupações de destino, e Z_i é a função linear de **X**.

$$m_{ij} = \frac{\partial P_i}{\partial X_j} = \left[\frac{\exp\ (Z_i)}{\sum_k \exp\ (Z_k)} \quad \frac{\partial Z_i}{\partial X_j} \right] - \left[\frac{\exp\ (Z_i)}{(\sum_k \exp\ (Z_k))^2} \Sigma_k\ (\exp\ (Z_k) \frac{\partial Z_k}{\partial X_j}) \right] = A$$

de forma que:

$$\varepsilon_{ij} \quad \frac{\partial P_i}{X_j} \quad \frac{X_j}{\partial P_i}$$

$$- X_j \frac{\sum_k \exp\ (Z_k)}{\exp\ (Z_i)} \ (A)$$

$$X_j \left[\frac{\partial Z_i}{\partial X_j} - \left(\frac{\exp\ (Z_i)}{\sum_k \exp\ (Z_k)} \quad \frac{\partial Z_k}{\partial X_j} + \frac{\exp\ (Z_m)}{\sum_k \exp\ (Z_k)} \quad \frac{\partial Z_m}{\partial X_j} \right) \right]$$

$$= X_j \left[\frac{\partial Z_i}{\partial X_j} - \sum_k\ (P_k \frac{\partial Z_k}{\partial X_j}) \right]$$

Dado que $\frac{\partial Z_i}{X_j} \quad \hat{\beta}^i_j; \ \ldots; \ \frac{\partial Z_m}{\partial X_j} = \hat{\beta}^i_j$

$$\varepsilon_{ij} - X_j \left[\hat{\beta}^i_j \quad \sum_k\ (p_k \hat{\beta}^k_j) \right]$$

TABELA 5

COEFICIENTES DE MODELOS DE PROBABILIDADES — MOBILIDADE A PARTIR DAS OCUPAÇÕES (URBANAS) HIERARQUICAMENTE INFERIORES (OCUPAÇÃO 10) — 1960-1965

Estabilidade na Ocupação 10	C	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
	2,081	0,020	0,014	0,010	0,161	−0,314	−0,172	−0,922	−0,732	−0,166
Mobilidade										
Ocupação 9	0,004	0,001	0,006	0,011	0,756[a]	−0,515[c]	−0,065	−0,741[b]	−0,193	(−0,52)
	(0,01)	(0,07)	(0,39)	(0,21)	(2,73)	(−1,78)	(0,27)	(−2,01)	(−0,58)	−0,067
Ocupação 8-7-6	−1,468[c]	−0,048[a]	0,019	−0,041	0,250	0,752[a]	0,195	1,993[a]	1,397[c]	0,320
	(−1,81)	(−3,03)	(1,12)	(−0,78)	(0,80)	(2,50)	(−0,78)	(2,77)	(1,99)	(1,15)
Ocupação 5-1	−0,618	0,028[b]	−0,040	0.20	−1,167[a]	0,077	0,492[d]	−0,330	−0,472	−0,087
	(1,11)	(1,99)	(−2,23)	(0,37)	(−2,85)	(0,30)	(1,56)	(−0,80)	(−1,23)	(−0,32)

OBS.: 1. Na primeira linha da tabela, X_1 experiência geral (no mercado de trabalho de Monterrey, unicamente); X_2 experiência profissional; X_3 educação; X_4 especialização; X_5 origem sócio-econômica (o pai não tinha uma profissão manual quando o entrevistado tinha 20 anos); X_6 = origem geográfica (não rural); X_7 = empregados em firmas com mais de 5, mas menos de 50 trabalhadores; X_8 o mesmo, com mais de 50 trabalhadores; X_9 = setor (trabalhadores em atividades "dinâmicas")

2. Na primeira coluna da tabela, ocupação 9 = ocupações manuais inferiores, ocupação 8-7-6 ocupações industriais ocupação 5-1 = ocupações não-manuais e administrativas.

3. Os números entre parênteses são valores 6 assintóticos.

4. As letras acima dos números indicam rejeição à hipótese nula (Ho: $\hat{\beta}x$ 0): nos seguintes níveis de significância (a) 1% ou mais; (b) = 5%; (c) = 10% = 10%; (d) = 20%.

A tabela 5 apresenta os coeficientes estimados para mobilidade ascendente a partir das ocupações hierarquicamente mais baixas no período 1960-65(50). Pode-se observar de imediato que, enquanto quase todas as variáveis mostravam-se estar, no período, relacionadas de forma significativa com a mobilidade, muitos dos valores dos parâmetros individuais não são estatisticamente diferentes de zero. Ou seja, os efeitos das variáveis individuais não têm a mesma importância para diferentes mudanças a partir de uma origem comum.

Ao associarmos os resultados da tabela 5 com as elasticidades da tabela 6, podemos ter uma idéia mais clara a respeito da contribuição dos parâmetros (ocupacionalmente específicos) das variáveis de experiência e escolaridade. No período em questão, a experiência geral contribuiu forte mas negativamente para a probabilidade de o indivíduo alcançar as profissões industriais; ela exerceu um efeito oposto sobre as mudanças para profissões não-manuais e administrativas mas, neste caso, o efeito quantitativo de uma experiência maior é altamente inelástico.

O impacto da experiência ocupacional foi surpreendentemente incerto e negligenciável, assim como o da educação — mais surpreendentemente ainda. Como indicam os resultados de probabilidade da tabela 7, doze anos a mais de escolaridade contribuíram para um aumento de apenas 15% na probabilidade de mudanças para profissões não-manuais e administrativas.

Por outro lado, o impacto das variáveis restantes foi decisivo, como pode ser observado na tabela 8. O grande aumento na probabilidade de mudança para ocupações industriais é, sem dúvida, o resultado do histórico seletivo e das características da empresa, enquanto a diminuição no número de mudanças para ocupações não-manuais e administrativas parece refletir o impacto altamente negativo da especialização anterior em outros tipos de trabalhos.

Estes resultados indicam a importância da especialização, do histórico e das características da empresa no processo de mobilidade, cuja complexidade e variabilidade freqüentemente frustram a interpretação dada pelo modelo de "capital humano" Na ver-

(50) Por motivos de espaço, não apresentamos aqui os resultados para as demais ocupações e períodos. O leitor interessado poderá encontrá-los em Vieira da Cunha (1980, capítulo 4).

TABELA 6

ELASTICIDADES DO MODELO DE PROBABILIDADE — MOBILIDADE A PARTIR DAS OCUPAÇÕES (URBANAS) HIERARQUICAMENTE INFERIORES (OCUPAÇÃO 10) — 1960-1965

Variável	Estável	Mobilidade para:		
		Ocupações manuais inferiores (Ocup. 9)	Ocupações Industriais (Ocup. 8-7-6)	Ocupações não-manuais e administrativas (Ocup. 5-1)
X_1	0,458	− 0,337	− 1,474	0,282
X_2	0,172	− 0,052	0,099	− 0,631
X_3	0,025	− 0,009	− 0,141	0,034

OBS.: Ver tabela 5. As elasticidades são avaliadas pelo valor médio das variáveis no ano de origem, isto é, 1960.

TABELA 7

VARIAÇÕES EM ESCOLARIDADE E PROBABILIDADES DE MUDANÇA OCUPACIONAL A PARTIR DAS OCUPAÇÕES (URBANAS) HIERARQUICAMENTE INFERIORES (OCUPAÇÃO 10) — 1960-1965

Escolaridade	Estáveis	Mobilidade para:		
		Ocup. manuais inferiores (Ocup. 9)	Ocupações Industriais (Ocup. 8-7-6)	Ocup. não-manuais e administrativas (Ocup. 5-1)
Nenhuma educação	81,60	11,00	3,72	3,68
1 ano	81,71	11,03	3,54	3,72
2 anos	81,81	11,06	3,36	3,76
Média	**81,88**	**11,08**	**3,25**	**3,80**
6 anos	82,14	11,16	2,75	3,94
12 anos	82,46	11,29	2,04	4,22

OBS.: As probabilidades são avaliadas pelos valores médios das variáveis restantes, medidas no ano de origem, isto é, 1960.

TABELA 8

PROBABILIDADE ESTIMADA DE MUDANÇA A PARTIR DAS OCUPAÇÕES (URBANAS) HIERARQUICAMENTE MAIS BAIXAS COM OU SEM EFEITOS DAS VARIÁVEIS BINARES — 1960-1965[a]

Variáveis	Estáveis	Mobilidade para:		
		Ocup. manuais inferiores (Ocup. 9)	Ocupações industriais	Ocup. manuais e administrativas (Ocup. 5-1)
Sem efeitos binares[b]	81,18	10,36	3,93	4,52
Com efeitos binares[c]	56,85	9,26	30,47	3,42

OBS.: (a) Estimadas pelos valores médios de experiência ocupacional e escolaridade. As variáveis binares incluem especialização, histórico sócio-econômico e geográfico, tamanho da empresa e setor.

(b) Exceto para a variável X_8 (na tabela 5) = 1.

(c) Exceto para a variável X_8 = 0.

dade, se tentássemos empregar o "modelo educacional" — mais simples e difundido — interpretando seus coeficientes como sendo o impacto direto e indireto do capital humano sobre a mobilidade, isto levar-nos-ia, inevitavelmente, a resultados enviesados(51). Isto é, no entanto, um debate à parte e que deve ser objeto de um trabalho específico. Nosso objetivo aqui foi explicitar uma abordagem metodológica alternativa de forma a permitir o estudo do *processo* de mobilidade ocupacional; portanto, as questões substantivas na interpretação dos resultados devem aguardar uma discussão mais completa.

Conclusão: Uma Sugestão Metodológica Final

Para finalizar, sugerimos aqui uma forma pela qual a metodologia anteriormente descrita pode ser empregada para abordar os problemas essenciais. Ao tratar da mobilidade, estamos, com efeito, analisando superposições de "cohorts" populacionais definidos pelo ano de entrada do trabalhador na força de trabalho, todos eles, salvo o último, envelhecendo ao aproximarmos do ano em que se realizou a coleta de dados. Eis aqui o problema. Com a idade, diminui a mobilidade e variam também quase todos os atributos pessoais. Conseqüentemente, varia a probabilidade de qualquer mudança ocupacional calculada a partir dos valores médios destes atributos.

Pode-se presumir que algumas das mudanças nos valores médios (tais como na composição de tamanhos de firmas ou até mesmo no nível de escolaridade dos trabalhadores) resultem de elementos exógenos ao modelo. Ainda assim, é forçoso observar que as probabilidades estimadas pelo modelo, quando baseadas em médias diferentes, não são estritamente comparáveis. Nesta situação, a alternativa que se apresenta (tendo obtido uma função de probabilidade para todos os períodos) é avaliar as probabilidades de mudança com um conjunto fixo de médias. Na verdade, podemos ir mais longe e, para cada dois períodos e uma determinada mudança ocupacional, decompor a probabilidade de mudança em duas partes.

Consideremos que $P_{ij} \mid {}^{xx}_{yy}$ (escrito para maior simplicidade P^{xx}_{yy}) signifique a probabilidade de mudança da ocupação *i* para a *j*, durante um determinado qüinqüênio, estimada por meio de médias (sobrescrito) de um ano de comparação *xx* e coeficientes (subscrito) de um período iniciado no ano *yy*. Usando esta notação, a diferença nos valores da probabilidade entre os períodos pode ser expressa como:

$$P^{yy}_{yy} - P^{xx}_{xx} = \left(P^{yy}_{yy} - P^{yy}_{xx} \right) + \left(P^{yy}_{xx} - P^{xx}_{xx} \right)$$

O primeiro termo à direita da expressão mede a diferença entre as probabilidades estimadas com diferentes conjuntos de coeficientes, mas com médias iguais às do período final. Sua magnitude reflete variações na contribuição de um conjunto "igual" de atributos assignados a uma mesma trajetória ocupacional mas em diferentes períodos; esta variação, vista desde a perspectiva do período final. A variação deve-se unicamen-

(51) Considere-se a seguinte comparação: primeiro, a probabilidade é estimada com o modelo complexo (ex.: como na tabela 5), este é, então, reestimado de acordo com o "modelo educacional" mais restrito (neste caso, incluindo unicamente as variáveis de experiência e escolaridade mais uma constante). Caso definamos a tendência (viés) como **b** = (R-T)/T, onde **R** é a probabilidade estimada por meio do modelo restrito e **T** é a probabilidade estimada pelo modelo total, os resultados para mudanças originais nas ocupações hierarquicamente inferiores (ocupação 10) no período 1960-65 são: (a) para os trabalhadores estáveis, + 8,73%; (b) para mudanças com destino nas ocupações manuais semiqualificadas, —40,16%; (c) para os destinos nas ocupações industriais, —84,31%; (d) para os destinos nas ocupações não-manuais e administrativas, +0,79%.

te a mudanças intertemporais no padrão de interação entre oferta e demanda e pode, por isso, ser qualificada como de natureza estrutural. Assim, o componente estrutural capta a diferença entre a probabilidade de mudança estimada para uma determinada trajetória e período e a probabilidade que o trabalhador "representativo" do período teria tido caso iniciasse a mesma nudança durante o ano de comparação.

O segundo termo à direita da expressão capta o efeito dos diferentes valores médios aplicados a um conjunto fixo de coeficientes, neste caso, os do período base ou de comparação. Como já foi definido, esta diferença é devida às mudanças intertemporais ocorridas na distribuição ocupacional das características dos trabalhadores. Dado que foram avaliadas para o período base mas vistas desde a perspectiva de um trabalhador com qualificações "médias" do período final, as diferenças indicam a vantagem (desvantagem) relativa deste trabalhador "vis-à-vis" seus colegas de gerações anteriores. Podemos denominar isto de *componente demográfica* na decomposição das diferenças de probabilidade entre os períodos.

No estudo em que se originou este artigo, fizemos estas decomposições e tentamos demonstrar que as variações intertemporais de mobilidade são influenciadas, principalmente, por mudanças na estrutura de demanda. As mudanças nas características dos trabalhadores, salvo aquelas diretamente ligadas ao envelhecimento dos "cohorts" (e que, no entanto, podem ser dificilmente considerados como submissos aos instrumentos da política econômica), têm um papel secundário e — o que é mais importante — no processo de mobilidade, elas desempenham um papel dependente das variações nas características estruturais(52).

(52) Ver Vieira da Cunha (1980, capítulos 5 e 6).

Referências Bibliográficas

BALÁN, J.; BROWNING, H. & Jelin, E. *Men in a developing society (geographic and social mobility in Monterrey, Mexico).* Austin, The University of Texas Press, 1973.

BARR, N. & HALL, R. The probability of dependency on public assistance. Unpublished manuscript. Cambridge, Mass., Massachusetts Institute of Technology, Department of Economics, 1973.

BEN-PORATH, Y. The production of human capital over time. *In:* HANSEN, W. Lee (ed.). *Education, income and human capital.* New York, NBER-Columbia University Press, 1970. p. 129-146.

BISHOP, Y.; FIENBERG, S. & HOLLAND, P. *Discrete multivariate analysis: theory and practice.* Cambridge, Mass., The MIT Press, 1975.

BLALOCK, H. M. *Social statistics.* 2 ed. New York, McGraw-Hill Book Co., 1972.

BLAU, P. & DUNCAN, O. *The american occupational structure.* New York, John Wiley and Sons, Inc., 1967.

BLUMEN, I.; KOGAN, M. & McCARTHY, P *The industrial mobility of labor as a probability process.* Ithaca, N.Y., Cornell Studies of Industrial and Labor Relations, 1955. v. 6.

BOCK, R.D. *Multivariate statistical methods in behavioral research.* New York, McGraw-Hill Book Co., 1975.

BOSKIN, M. A conditional logit model of occupational choice. *Journal of Political Economy. 82:* 389-398, 1974.

COX, D. R. *Analysis of binary data.* London, Methuen and Co. Ltd., 1970.

DAVIS, J. A. Hierarchical models for significance tests in multivariate Contingency tables: an exegesis of Goodman's

recent papers. *In:* COSTNER, H. L. (ed.) *Sociological methodology: 1973--1974.* San Francisco, Jossey-Bass Publ., 1974. p. 189-231.

DUNCAN, O. D. Methodological issues in the analysis of social mobility. In: SMELSER, N. & LIPSET, S. (ed.) *Social structure and mobility in economic development.* Chicago, Aldine Publ., 1966. p. 51-97.

FIENBERG, S. *The analysis of cross-classified categorical data.* Cambridge, Mass., The MIT Press, 1977.

GINSBERG, R. Critique of probabilistic models: application of the Semi-Markov model to migration. *Journal of Mathematical Sociology. 2:* 63-82, 1972.

GOODMAN, L. A. How to ransack social mobility tables and other kinds of cross-classification tables. *American Journal of Sociology. 75:* 1-40, 1969.

———— The multivariate analysis of qualitative data: interactions among multiple classifications. *Journal of the American Statistical Association. 65:* 226-256, 1970.

HABERMAN, S. *The analysis of frequency data.* Chicago, University of Chicago Press, 1974.

HALEY, W. Estimation of the earnings profile from optimal human capital accumulation. *Econometrica. 44:* 1223-1238, 1976.

HAUSMAN, J. & WISE, D. A conditional probit model for qualitative choice: discrete decisions recognizing interdependence and heterogeneous preferences. *Econometrica. 46:* 403-426, 1978.

HECKMAN, J. A life-cycle model of earnings, learning, and consumption. *Journal of Political Economy. 84:* S11-S4, 1976.

HODGE, R. W. Occupational mobility as a probability process. *Demography. 3:* 19-34, 1966.

HOLLIS, M. & NELL, E. *Rational economic man.* Cambridge, The University Press, 1975.

KOHN, M. Beyond regression: a guide to conditional probability models in econometrics. Unpublished manuscript. Hebrew University, Department of Economics, 1976.

KOHN, M.; MASNSKI, C. & MUNDEL, D. An empirical investigation of factors which influence college-going behavior. *Annals of Economic and Social Measurement. 5:* 391-419 1976.

LILLARD, L. & WILLIS, R. Dynamic aspects of earning mobility. *Econometrica. 46:* 985-1012, 1978.

MAcCALL, J. A theory of income mobility, racial discrimination and economic growth. *Income mobility racial discrimination and economic growth.* Lexington, Lexington-Books, 1973.

MAcRAE, E. Estimation of time-varying Markov processes with aggregate data. *Econometrica. 45:* 183-198, 1977.

MANSKI, C. & LERMAN, S. The estimation of choice probabilities from choice based samples. *Econometrica. 45:* 1977-1988, 1977.

McFADDEN, D. Quantal choice analysis: a survey. *Annals of Economic and Social Measurement. 5:* 363-390, 1976a.

———— A comment on discriminant analysis 'versus' logit analysis. *Annals of Economic and Social Measurement. 5:* 511-524, 1976b.

———— Conditional logit analysis of qualitative choice behavior. Institute of Urban and Regional Develop-

ment, University of California (Berkeley), Working Paper 199/BART 10. Berkeley, Califórnia. Também *In:* ZAREMBKA, P. (ed.). *Frontiers in econometrics.* New York, Academic Press, 1973.

McFARLAND, D. Intragenerational social mobility as a markovian process: including a time stationary markovian model that explains observed declines in mobility rates over time. *American Sociological Review. 35:* 463-76, 1970.

NERLOVE, M. & PRESS, S. Review of discrete multivariate analysis: theory and practice. Unpublished manuscript. Northwestern University, Department of Economics, 1977.

———— Multivariate log-linear probability models for the analysis of qualitative data center for statistics and probability. Discussion paper n.o 1. Evanston, I11., Northwestern University, 1076.

———— Univariate and multivariate log-linear and logistic models. Rand Corporation technical Report R-1306-EDA/NIH. Santa Monica, California, 1973.

PALMER, G. *Labor mobility in six cities.* New York, Social Science Research Council, 1954.

PASTORE, J. Desigualdade e mobilidade social no Brasil. 1979.

PRAIS, S. Measuring social mobility. *Journal of the Royal Statistical Society.* Series A, *18:* 56-66, 1955.

PRESS, S. J. & WILSON, S. Choosing between logistic regression and discrimination analysis. *Journal of the American Statistical Association. 73:* 699-705, 1978.

REYNOLDS, H. T. *The analysis of cross-classification.* New York, The Free Press, 1977.

ROSEN, S.A. theory of life earnings. *Journal of Political Economy. 84:* S45-S67, 1976.

SCHMIDT, P. & STRAUSS, R. The prediction of occupation using multiple logit models. *International Economic Review.* 16: 471-486, 1975.

SEWELL, W. & HAUSER, R. *Education, occupation and earnings.* New York, Academic Press, Inc., 1975.

SHORROCKS, A. Income mobility and the Markov assumption. *Economic Journal. 86:* 566-78, 1976.

SIMPSON, E. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society.* B, *12:* 238-241, 1951.

SINGER, B. & SPILERMAN, S. Some methodological issues in the analysis of longitudinal surveys. *Annals of Economic and Social Measurement. 5:* 447-474, 1976.

———— Social mobility models for heterogeneous populations. *In:* COSTNER, H. L. (ed.) *Sociological methodology 1973-1974.* San Francisco, Jossey-Bass Pub., 1974.

SORENSEN, A. Models of social mobility. *Social Science Research. 4:* 65-92, 1975.

SPILERMAN, S. The analysis of mobility processes by the introduction of independent variables into a Markov chain. *American Sociological Review. 37:* 277-294, 1972.

STEWMAN, S. Markov models of occupational mobility: theoretical development and empirical support. Part 1: Careers. *Journal of Mathematical Sociology. 4:* 201-245, 1976.

THEIL, H. *Statistical decomposition analysis.* Amsterdam, North-Holland Pub. Co., 1972.

———————— On the estimation of relationships involving qualitative variables. *American Journal of Sociology. 76:* 103-154, 1970.

UPTON, G. Log-linear models, screening and regional industrial surveys. *Regional Studies. 15:* 33-45, 1981.

VIEIRA DA CUNHA, P. Occupational mobility and labor market segmentation. Unpublished manuscript. University of California, Berkeley, Department of Economics, 1975.

———————— Structures of production and employment. Occupational change in Monterrey, Mexico: 1940-1965. Tese de doutoramento (inédita) apresentada à Universidade da Califórnia, Berkeley, 1980.

WISE, D. Personal attributes, job performance, and probability of promotion. *Econometrica. 43:* 913-931, 1975.