

**PENÉLOPE**

**QUETZAL EDITORES  
LISBOA, 1988**

Composição e impressão: *Tipografia Guerra, Viseu*  
Capa e arranjo gráfico: *Rogério Petinga*  
Depósito legal n.º 19712  
NQZ.09.001.37.88

ESTA contribuição para um espaço de apresentação e discussão de métodos de investigação em história pretende realçar as virtualidades da técnica de construção de modelos recursivos de causalidade para o teste de hipóteses e para a elaboração de modelos interpretativos das evidências empíricas observadas. Esta dupla função de *validação* e de *configuração* de hipóteses interpretativas constitui um forte incentivo para tornar a *path analysis* (nome pelo qual aquele método estatístico é usualmente designado) como um auxiliar precioso na investigação histórica.

A abordagem aqui feita não tem a intenção de proceder a um tratamento exaustivo deste tema. O objectivo fundamental é o de referir as características da *path analysis*, os problemas a que pode ser aplicada e a bibliografia existente sobre este tema. Para além das limitações inerentes à necessidade de apresentar uma visão sintética sobre aquela técnica estatística, opta-se por uma abordagem «de utilizador para utilizador», em que os objectivos de divulgação se sobrepõem à exaustão analítica sobre o tema.

Assim, este artigo desdobrar-se-á em duas partes essenciais:

- 1 — Num primeiro momento proceder-se-á à apresentação e discussão do tipo de problemas a que esta técnica se adequa.
- 2 — Em seguida será abordado o método de cálculo e de interpretação dos coeficientes de dependência que constituem o elemento fundamental dos modelos recursivos de causalidade.

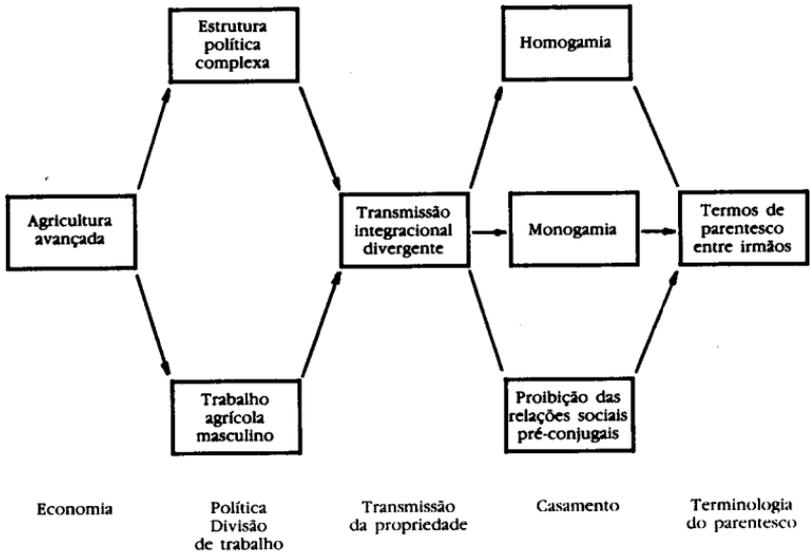
Esta exposição pressupõe por parte do leitor um conhecimento elementar sobre a teoria e o vocabulário da análise estatística subjacente à construção e interpretação de equações de regressão, embora sejam fornecidos elementos bibliográficos que permitem superar esta barreira.

## II

O recurso a métodos estatísticos e a um processo de análise quase-experimental<sup>1</sup> é hoje uma situação usual num número significativo de investigações históricas. O estabelecimento de modelos explicativos baseados na detecção de nexos causais entre um conjunto de variáveis é um elemento importante no processo de explicação, a par de outros tipos de explicação histórica<sup>2</sup>. Por vezes este aspecto não é considerado, quer devido a um posicionamento teórico segundo o qual as relações entre variáveis respeitantes às condutas humanas devem ser analisadas principalmente em termos de covariação, e não de relações de causalidade<sup>3</sup>, quer simplesmente porque se tende a iludir esta questão através da sua omissão das preocupações metodológicas do historiador. No entanto, o vocabulário utilizado para descrever e interpretar os processos sociais está cheio de palavras em que ressalta a sua conotação com um modelo de explicação causal. Fala-se de «factores cruciais», de «elemento dependente» ou de «aspectos preponderantes», sem que se dê conta de que se tratam de vocábulos em que está presente uma conotação causal.

Por outro lado, não se trata apenas de assumir que para a explicação do comportamento de um determinado fenómeno se conjugam um conjunto de factores. Uma atitude deste tipo poderá traduzir-se no arrolamento de um conjunto de «causas», sem que se estabeleça o peso específico de cada uma delas, a sua contribuição relativa para a explicação. Tal como se tem revelado de reduzida utilidade o estabelecimento de relações unicasais, também a mera enunciação de múltiplas relações de causalidade se transforma numa operação de reduzida eficácia analítica. Daí que surja como potencialmente útil a adopção de métodos estatísticos que permitam testar a existência de nexos causais entre as diferentes variáveis e detectar a sua importância relativa num modelo explicativo.

Fig. 1 — Modelo causal: implicações sociais da agricultura avançada (Goody, 1976: 38)

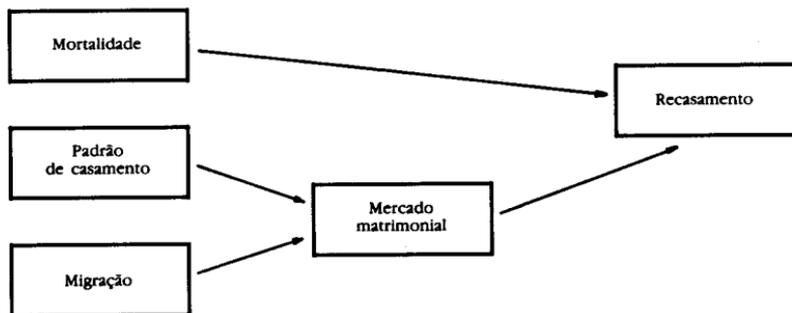


Tomemos um exemplo bem preciso. Num dos capítulos de *Production and Reproduction*, Jack Goody aborda a necessidade de passar de uma situação em que detectou associações entre as transformações no sistema produtivo de uma sociedade e as mudanças nas instituições do parentesco<sup>4</sup>, para a explicação das causalidades que presidem às correlações encontradas. Para tal, vai elaborar um modelo causal que inter-relacione todas as variáveis de que dispõe e que podem explicar a terminologia do parentesco — sobre a organização política e económica de uma dada sociedade, sobre o sistema de transmissão de recursos produtivos entre gerações e sobre o sistema de casamento. Num primeiro momento, este modelo causal surge como um conjunto de hipóteses que tentam explicar as inter-relações entre as diferentes variáveis. Como pode ser visto através da fig. 1, trata-se de um diagrama que postula a existência de relações

de causalidade unidireccionais entre um conjunto de variáveis. A sua diferença face à existência de associações de covariação entre cada par de variáveis, mensuráveis através de coeficientes de correlação<sup>5</sup>, reside em dois aspectos fundamentais: em primeiro lugar, associa-se uma relação de causalidade à covariação observada entre duas variáveis, definindo que esse nexos causal se processa num só sentido; em segundo lugar, não são consideradas todas as relações entre cada par de variáveis, assumindo-se a existência de laços causais apenas entre algumas de entre elas. A adequação do modelo causal esboçado às relações existentes entre as diferenças variáveis vai ser então testada através da *path analysis*, que possibilitará igualmente o estabelecimento do peso que pode ser atribuído a cada variável independente na explicação de uma variável dependente<sup>6</sup>.

Numa investigação desenvolvida sobre as práticas de recasamento em Portugal durante o século XIX<sup>7</sup>, utilizei também a construção de um modelo causal que traduzisse a explicação que variáveis de índole demográfica conseguem trazer à frequência do recasamento. A fig. 2 apresenta a configuração desse modelo, que procurei testar recorrendo igualmente à técnica da *path analysis*.

Fig. 2 — Modelo causal do recasamento (Silva, 1986)



Para além de possibilitar a elucidação sobre a adequação de um modelo de análise às relações observáveis entre um conjunto de variáveis, a *path analysis* utiliza métodos estatísticos que têm sido objecto de um prolongado estudo sobre a sua congruência e validade, o que se torna num elemento de segurança para a sua utilização criteriosa. Como se poderá ver mais adiante, o cálculo de equações de regressão é a técnica subjacente à *path analysis*, existindo todo um conjunto de pressupostos que devem ser cumpridos e testados para se poderem ter resultados susceptíveis de adquirirem um significado estatístico.

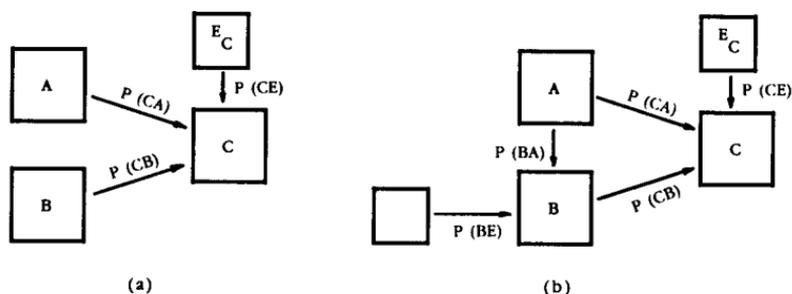
## III

A explicitação dum modelo numa das suas formas mais simples é exemplificada na fig. 3 (a). Traduzida através de uma equação de regressão múltipla, esta teria a seguinte forma:

$$C_i = \alpha + \beta_1 A_i + \beta_2 B_i + \epsilon$$

O valor atribuível ao efeito directo entre A e C, simbolizado por P (CA), não é mais do que a standardização do coeficiente de regressão  $\beta_1$ . Sendo A e B as duas variáveis independentes face à variável dependente C, o coeficiente de regressão standardizado P (CA) seria interpretado como a intensidade da variação produzida na variável dependente C, devida a uma variação standardizada na variável independente A, mantendo-se constante a variável independente B<sup>8</sup>. Os coeficientes de regressão standardizados — também designados como coeficientes de dependência, *beta weights* ou *path coefficients* — podem ser considerados como o *efeito directo* da variável independente a que se referem sobre a variável dependente.

Fig. 3 — Dois tipos de estrutura causal



Por outro lado, o termo da equação de regressão  $\epsilon$ , muitas vezes designado como o factor de «erro», é a diferença entre o valor observado para a variável dependente e o valor previsto pela equação de regressão das variáveis independentes. Trata-se pois de um termo de «perturbação» ou «residual» que recolhe os efeitos de outras variáveis não contempladas no modelo (ou porque não foram ainda incorporadas no modelo causal ou porque não são directamente mensuráveis) e os efeitos devidos a variações aleatórias. O coeficiente de dependência entre os efeitos residuais da equação de regressão e a variável dependente C é obtido através da seguinte fórmula:

$$P(CE) = \sqrt{1 - R^2}$$

A quantidade  $\sqrt{1 - R^2}$ , por vezes apelidada de coeficiente de alienação, representa a raiz quadrada da proporção da variação na variável dependente não explicada pela equação de regressão. A sua integração no modelo causal estabelece o efeito directo dos elementos residuais na variação de C, dando conta do peso relativo de factores aleatórios ou não integrados na equação de regressão.

A fig. 3 (b) representa uma complexificação do modelo causal baseado numa única equação de regressão. Neste caso, para

além do efeito directo entre A e C, existe um efeito indirecto mediatizado por B, através do laço causal entre A e B. Isto é, B assume-se como uma variável independente face a C, mas dependente face a A. Neste caso, torna-se necessário calcular duas equações de regressão. Os coeficientes de regressão estandardizados permitem achar os efeitos directos de cada variável independente, sendo necessário formular tantas equações de regressão quantas as variáveis sucessivamente dependentes que fazem parte do modelo. No caso do diagrama causal referente ao recasamento, tal como é formulado na fig. 2, seriam necessárias duas equações de regressão múltipla — o mercado matrimonial seria variável dependente numa das equações de regressão e variável explicativa numa outra.

Uma primeira tarefa prévia ao cálculo dos coeficientes de dependência é, pois, a formalização num diagrama das hipóteses explicativas das relações entre as diferentes variáveis. Trata-se, sem dúvida, da etapa mais difícil da *path analysis*<sup>9</sup>. Pressupõe um trabalho de escolha dos melhores indicadores para as variáveis que se pretendem analisar, bem como uma arquitectura clara das hipóteses teóricas que orientam a investigação. Nunca é de mais salientar a necessidade de uma integração óptima entre teoria, fontes e metodologia de investigação. Deficiências na apreciação da qualidade das fontes e na sua crítica, falhas na recolha de indicadores adequados a variáveis que não são directamente mensuráveis (por lacunas nos dados ou por serem variáveis latentes), incompreensão dos pressupostos invioláveis inerentes ao uso de determinadas técnicas estatísticas podem levar qualquer análise — não importa quão sofisticada — a exercícios arbitrários de manipulação numérica.

A utilização da regressão múltipla como método de análise estatística contém, nomeadamente, um conjunto de pressupostos que não são passíveis de serem escamoteados. Prendem-se, acima de tudo, com a necessidade de os resíduos simbolizados por  $\epsilon$  na equação de regressão serem uma variável *estocástica* (a sua variação apenas depende de factores aleatórios) e de esta-

rem sujeitos à hipótese de independência entre os diferentes valores dos resíduos<sup>10</sup>. Existe um conjunto de testes à existência de alguma situação em que qualquer destes pressupostos seja violado, podendo inclusivamente corrigir-se os efeitos de autocorrelação relativamente ao último caso. Quanto ao carácter estocástico que devem assumir os resíduos, a importância deste aspecto prende-se com a preocupação em esgotar as variáveis explicativas. Isto é, se a distribuição do valor dos resíduos não é puramente aleatória, existindo, pelo contrário, uma covariação entre os valores de  $\epsilon$  e os valores de qualquer uma das variáveis independentes, então é necessário proceder à identificação das variáveis não integradas na equação de molde a também as considerar.

Contudo, não são apenas os problemas relativos à necessidade de esgotar a capacidade explicativa do modelo causal, através de um apuramento exaustivo das variáveis relevantes, que podem ser encontrados. Ao associarmos um nexos causal entre duas variáveis, podemos estar a estabelecer uma relação *espúria*. Um nexos causal directo só pode ser estabelecido quando não existe entre uma variável independente e uma variável dependente outro factor que «eclipse» aquele efeito directo — empregando a terminologia de David Papineau num estudo sobre a interacção entre a análise estatística e a formulação de conclusões sobre relações causais<sup>11</sup>.

Um exemplo citado por Blalock permite transmitir a ideia do que é uma relação espúria. Qual a relação causal associada à existência de uma correlação negativa entre o consumo de gelados e as taxas de delinquência juvenil? Poder-se-ia pensar que o consumo de gelados teria um efeito de impedir a delinquência. Ou, ao invés, que a prática da delinquência faria desaparecer nos jovens o seu gosto pelos gelados... Blalock conclui que qualquer das interpretações seria rapidamente rejeitável como absurda, mas que conclusões igualmente absurdas tinham já sido consideradas como correctas<sup>12</sup>. Se se tivesse introduzido uma outra variável como o rendimento, então chegar-se-ia à con-

clusão que era esta última variável que estava a afectar a variação de qualquer uma das outras duas variáveis, de tal forma que era obtida uma correlação negativa entre elas. Uma correlação parcial entre a delinquência juvenil e o consumo de gelados, controlando os efeitos do rendimento, seria reduzida a uma expressão próxima de zero, o que comprovaria a ideia de que a covariação entre as duas variáveis era completamente espúria.

O estabelecimento de uma rede de efeitos directos e indirectos a partir de um diagrama causal pode incorrer num conjunto de erros de formulação de relações espúrias entre variáveis. A concretização dos testes à não violação dos pressupostos subjacentes à regressão pode desde logo fornecer uma primeira detecção de erros deste tipo. A existência de uma covariação entre os valores dos resíduos e uma variável independente pode ser sintomático de que existe uma variável oculta agindo simultaneamente sobre a variável «independente» e a variável dependente.

Por outro lado, como foi dito anteriormente, o estabelecimento de um diagrama causal pressupõe uma «economia» nos efeitos directos considerados, não sendo formuladas relações causais directas entre todas as variáveis. Ora a inexistência de covariação ou a assunção de que essa covariação é espúria e

---

### Quadro 1

Correlações parciais para teste do modelo da fig. 4

---

$$R_{34.1} = -0.098 \quad (s = 0.057)$$

$$R_{35.1} = 0.023 \quad (s = 0.354)$$

$$R_{16.2345} = -0.001 \quad (s = 0.597)$$

---

Notas — Forma de leitura das correlações parciais: correlação da variável 3 com a variável 4, controlando os efeitos da variável 1.

— s: nível de significância.

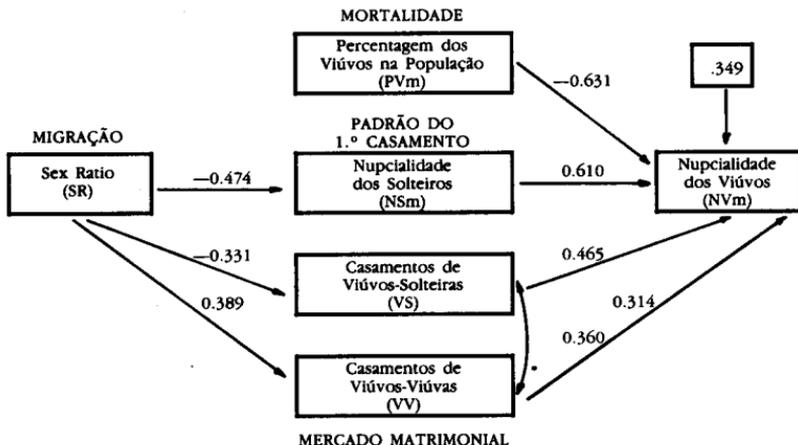
---

não causal implica necessariamente que se teste essa hipótese. Para tal utiliza-se o mesmo método de cálculo de correlações parciais, no pressuposto de que as correlações parciais entre duas variáveis para as quais não se prevê qualquer relação de causalidade apresentam valores tendendo para zero e com níveis de significância baixos.

A fig. 4 representa o diagrama causal do recasamento da população masculina, que apresenta algumas diferenças face ao diagrama que tinha sido formulado como correspondendo ao feixe de hipóteses inicial<sup>13</sup>. Nele se detectam algumas variáveis para as quais não é apresentada qualquer relação causal e que são testadas através de um conjunto de correlações parciais exemplificadas no quadro 1. Como se pode constatar, os valores das correlações parciais alcançam níveis diminutos ou praticamente nulos.

O cálculo dos coeficientes de dependência ou *beta weights* não traz qualquer dificuldade, sendo apresentados nos progra-

Fig. 4 — Modelo causal do recasamento dos viúvos



- Notas: X Y — efeito causal directo  
 X Y — inexistência de covariação ou covariação espúria e não causal  
 X Y — covariação não analisada ou em que existem dúvidas quanto á direcção causal

mas estatísticos de regressão múltipla com a designação de «beta» (como no caso do SPSS) ou de «b estandardizado» (BIOM). No entanto, quando se dispõe unicamente de variáveis ordinais e nominais, o cálculo dos coeficientes de dependência através da equação de regressão múltipla está comprometido. O trabalho de Jack Goody atrás referido deparou com este problema, tendo-o ele solucionado pela dicotomização das variáveis nominais e pelo emprego do cálculo do coeficiente de associação  $\phi$  ( $\Phi$ ) como correspondente ao coeficiente de correlação para o nível de medida por intervalos<sup>14</sup>. Desde a data em que Goody elaborou o seu estudo deram-se, todavia, alguns passos importantes na adequação de métodos de análise de relações causais a variáveis de tipo qualitativo. A utilização de modelos lineares logarítmicos (*log-linear models*) a partir de quadros de contingências tem sido uma das áreas mais recentemente desenvolvidas para lidar com as características próprias das variáveis no âmbito das ciências sociais<sup>15</sup>.

A interpretação dos coeficientes de dependência associados com cada relação causal é igualmente simples. O seu cálculo pela estandardização dos coeficientes  $\beta$  da regressão fá-los variar numa escala possível entre  $-1$  e  $+1$ , assumindo cada variável explicativa ou independente o seu peso relativo na variação da variável dependente. A título de exemplo, o modelo causal apresentado quanto ao recasamento masculino associa — ao contrário do que usualmente é pretendido — um efeito directo negativo à mortalidade<sup>16</sup>. Tradicionalmente tem-se entendido que quanto maior é a mortalidade da população feminina casada maior é a taxa de recasamento masculino. No entanto, a efectiva criação de uma oportunidade para o recasamento devido à dissolução da união conjugal pela morte de um dos cônjuges não implica a associação de uma relação causal positiva oriunda daquele factor. A confusão entre oportunidade e probabilidade estatística gerou este equívoco. Por seu turno, a variável que em seguida detém o efeito directo mais elevado sobre a nupcialidade dos viúvos é o indicador relativo ao padrão

do primeiro casamento. Neste caso, as virtualidades do modelo causal já não são apenas as de desfazer um equívoco historiográfico, mas salienta-se uma conclusão inteiramente nova e que não seria passível de ser detectada sem a elaboração de uma estratégia de investigação deste tipo — a intensidade do recasamento não deve ser desligada dos modelos sociais e culturais que determinam o primeiro casamento, surgindo quase como que um seu epifenómeno.

Por outro lado, é possível decompor a covariação entre duas variáveis relacionadas por umnexo causal em efeitos directos (coeficientes de dependência) e em efeitos indirectos<sup>17</sup>, o que permite complexificar a análise, detectando-se nomeadamente quais as variáveis que, veiculando os seus efeitos unicamente através da relação causal directa, podem ser interpretadas como tendo uma relação mais constante na variação da variável dependente.

#### IV

O problema que habitualmente se coloca quando são apresentadas técnicas estatísticas como as que se referem à construção de modelos de causalidade diz respeito à sua adequação às fontes a partir das quais o historiador constrói os seus indicadores ou aos problemas que orientam a sua investigação. As diferenças entre investigação experimental (para a qual muitos dos métodos estatísticos foram inicialmente idealizados) e investigação não experimental, tal como a necessidade de solucionar os problemas de validade interna e externa das conclusões oriundas da utilização de métodos estatísticos<sup>18</sup>, exige o maior cuidado na adequação entre teoria, evidências empíricas e metodologia estatística, como salientei no início deste texto.

Por outro lado, muito raramente se aborda em termos correctos o problema do *isomorfismo* entre os métodos quantitativos utilizados e as hipóteses teóricas e analíticas subjacentes

à investigação. Esta é uma questão crucial que deveria orientar um profundo debate sobre o grau de adequação entre a lógica das operações estatísticas e a lógica das relações conceptuais<sup>19</sup>.

A apresentação aqui feita de dois estudos com problemáticas, contextos disciplinares e níveis de medida das variáveis tão diferentes pretendia chamar a atenção para a adequação da *path analysis* a situações radicalmente distintas. A ideia de que o método de prova e de raciocínio estatístico é apenas aplicável a variáveis que tomam a forma de intervalos ou a contextos disciplinares e analíticos específicos (como a história económica ou a demografia histórica) é um tremendo equívoco. A forma como Jack Goody finaliza a abordagem das relações entre economia, instituições e a terminologia do parentesco parece-me uma síntese feliz sobre a utilidade deste tipo de métodos<sup>20</sup>:

«While it is important not to overvalue the use of such techniques, the methods employed represent an advance that helps us to get a little beyond the circularity of structural functionalism and the much simpler unilineal, single-factor hypotheses that dog so much work in the social sciences. [...] They are still far too crude, but at least such methods tend to reduce the arbitrary element in our assumptions and therefore to replace dispute by argument about theories which can be tested and improved rather than simply rejected in favour of a different set of assertions.»

### NOTAS

<sup>1</sup> Sobre a relação entre processos de análise experimentais e quase-experimentais veja-se Blalock (1968), Cook e Campbell (1979).

<sup>2</sup> Sobre os diferentes tipos de explicação histórica veja-se Topolsky (1976: 536)

<sup>3</sup> Goody (1976) critica precisamente o funcionalismo estrutural de Radcliffe-Brown por optar por uma explicação deste tipo.

<sup>4</sup> Goody (1976; 23-40). Cf. também Goody *et al.* (1980: 33-47).

<sup>5</sup> Goody (1976-1980) utiliza o coeficiente de associação *pbi* ( $\phi$ ) para substituir o coeficiente de correlação.

<sup>6</sup> Sobre o conceito estatístico de independência veja-se Blalock (1979: 121)

<sup>7</sup> Cf. Silva (1986). Este trabalho fez parte de uma investigação comparada sobre a prática e a distribuição regional do recasamento na Península Ibérica desenvolvida no Núcleo de Sociologia Histórica do Instituto Gulbenkian de Ciências.

<sup>8</sup> O coeficiente de regressão  $\beta$  representa, por seu turno, a variação esperada na variável dependente C, devida à variação de uma unidade na variável independente A, mantendo-se B constante. Sobre a teoria da regressão linear veja-se Blalock (1979), Hanusheck *et al.* (1977) como introduções ao tema. Sobre a *path analysis*: Asher (1976), Blalock (1971; 1979: 469-482), Boudon (1971), Satorra *et al.* (1984). O significado do coeficiente de dependência ou *beta weight* pode ser melhor entendido se nos lembrarmos das relações entre o coeficiente de correlação (R) e o coeficiente de regressão ( $\beta$ ) no caso da regressão simples (cf. Blalock, 1979: 398-406).

<sup>9</sup> Cf. Mitchell (1980: 6-7)

<sup>10</sup> Sobre os pressupostos da regressão linear consulte-se a bibliografia referida na nota 8.

<sup>11</sup> Cf. Papineau (1978).

<sup>12</sup> Veja-se Blalock (1979: 469-470).

<sup>13</sup> Compare-se com o diagrama causal da fig. 2.

<sup>14</sup> Goody (1980: 44-46).

<sup>15</sup> Está fora do âmbito desta breve exposição uma abordagem desta metodologia, tal como não se referiu o método de cálculo dos coeficientes de dependência quando existem relações de causalidade recíproca entre as variáveis independentes e dependentes. A existência de efeitos de recorrência implica o abandono da metodologia subjacente aos modelos de causalidade recursivos aqui sintetizada. A bibliografia apresentada ao longo do texto possibilitará um contacto do leitor interessado com estes problemas e, simultaneamente, um aprofundamento dos aspectos aqui sumariados. Cf. Sanchez Carrión (1984b; 1984c) sobre os modelos lineares logarítmicos.

<sup>16</sup> A intensidade do efeito causal atribuível à mortalidade está, no entanto, inflacionada devido ao facto de se tratar de um indicador indirecto da mortalidade diferencial adulta entre regiões e consoante os sexos dos indivíduos casados.

<sup>17</sup> Para uma explicação acessível do seu cálculo veja-se Boudon (1971) e Blalock (1979: 477-482).

<sup>18</sup> Campbell e Cook (1979: 39-81).

<sup>19</sup> Veja-se a interessante abordagem deste aspecto por Mitchell (1980: 21-25) e Garbett (1980: 224-227). Peter Gould (1979), embora sob outra perspectiva, debruça-se igualmente sobre este problema.

<sup>20</sup> Goody (1976: 39-40).