Rev. Saúde públ., S. Paulo, 19:263-9, 1985.

ANÁLISE MULTIVARIADA; UM EXEMPLO USANDO MODELO LOG-LINEAR

José Maria Pacheco de Souza*
Maria Helena D'Aquino Benicio**

SOUZA, J.M.P. de & BENÍCIO, M.H.D'A. Análise multivariada: um exemplo usando modelo log-linear. Rev. Saúde públ., S. Paulo, 19:263-9, 1985.

RESUMO: Apresenta-se de forma resumida análise multivariada de dados categóricos, usando modelo log-linear para a situação de uma tabela de contingência 2 x 2 x 2.

UNITERMOS: Análise multivariada. Modelo log-linear.

INTRODUÇÃO

A técnica de análise multivariada de dados categóricos, mediante modelos log-linear ou modelo logito, é bastante útil em trabalhos na área de Saúde Pública e Epidemiologia, onde é comum se ter tabelas de contingência complexas, com grande número de variáveis.

O objetivo do presente trabalho é apresentar de forma resumida tal técnica, para a situação particular de três variáveis, cada uma com duas categorias mutuamente exclusivas, ou seja, para a situação de uma tabela de contingência $2 \times 2 \times 2$. Worcester [5] apresenta trabalho em linha semelhante, analisando também a situação para tabela 2×2. Vitaliano [6] analisa situação mais complexa em um estudo caso-controle.

O exemplo numérico consta de dados referentes ao trabalho de Benício [1]; um programa de computador que executa os algoritmos necessários à análise - ECTA -, escrito por Leo Goodman, encontra-se à disposição no Centro de Computação Eletrônica da Universidade de S. Paulo. Sobre o assunto há vários textos, de vários níveis de complexidade matemática [2,3,4].

MODELO LOG-LINEAR

Seja a distribuição teórica de frequências da Tabela 1, onde F_{ijk} é a freqüência teórica dos níveis i, j, k, respectivamente das variáveis 1, 2 e 3, onde i, j, k variam de 1 a 2. F_{112} é o número esperado teórico de indivíduos com a característica 1 da variável 1, com a característica 1 da variável 2 e com a característica 2 da variável 3.

Tomando-se logarítmo natural ln (base e = 2,71828...), pode-se demonstrar [2,3] que

$$\ln F_{ijk} = B + B_1(i) + B_2(j) + B_3(k) + B_{12}(ij) + B_{13}(ik) + B_{23}(kj) + B_{123}(ijk)$$

onde os B's são parâmetros que representam "efeitos", a exemplo da análise de variância;

$$B = \frac{1}{8} \sum_i \sum_j \sum_k \ln F_{ijk}$$

$$B_1(1) = \left(\frac{1}{4} \sum_j \sum_k \ln F_{1jk}\right) - B$$

* Do Departamento de Epidemiologia da Faculdade de Saúde Pública da Universidade de São Paulo — Av. Dr. Arnaldo, 715 — 01255 — São Paulo, SP — Brasil.
** Do Departamento de Nutrição da Faculdade de Saúde Pública da Universidade de São Paulo — Av. Dr. Arnaldo, 715 — 01255 — São Paulo, SP — Brasil.

$$B_2(1) = \left(\frac{1}{4} \sum_i \sum_k \ln F_{i1k}\right) - B \qquad B_{12}(11) = \left(\frac{1}{2} \sum_k \ln F_{11k}\right) - B_1(1) - B_2(1) + B, \text{ etc.}$$

TABELA 1

Distribuição teórica de freqüências para três variáveis, cada uma com duas categorias. A variável 1 eventualmente pode ser considerada "resposta" e sua categoria 2 evento desfavorável.

Variável 1 (resposta)	Variável 2	Variável 3			
		Categoria 1		Categoria 2	
		Categoria 1	Categoria 2	Categoria 1	Categoria 2
Categoria 1		F_{111}	F_{121}	F_{112}	F_{122}
Categoria 2 (evento desfavorável)		F_{211}	F_{221}	F_{212}	F_{222}

Note-se que B é a média aritmética dos logarítmos naturais das freqüências teóricas; $B_1(1)$ mede o desvio da média aritmética dos logarítmos das freqüências teóricas da categoria 1 da variável 1 em relação à média geral B, ou seja, mede o "efeito" 1 da variável 1; analogamente tem-se $B_1(2)$, $B_2(1)$, $B_2(2)$, etc., sendo $B_1(1) + B_1(2) = 0$; $B_2(1) + B_2(2) = 0$; $B_3(1) + B_3(2) = 0$, etc.

Os parâmetros com subscrito duplo e triplo são os mais importantes para a análise, sendo aqueles que medem as possíveis associações (interações) entre variáveis. Assim, $B_{12}(11)$ é o parâmetro que indica se as categorias 1 da variável 1 e 1 da variável 2 estão associadas; se $B_{12}(11) = 0$ não há associação; se $B_{12}(11) < 0$ tem-se associação negativa; se $B_{12}(11) > 0$ tem-se associação positiva.

As freqüências F e os parâmetros B são desconhecidos. A partir de modelos e das freqüências observadas f_{ijk}, obtém-se estimativas das freqüências F_{ijk} e dos B's, denotados, respectivamente, por E_{ijk} e b.

AJUSTE E TESTE DE MODELO: TESTE DE B

Considera-se modelo adequado para descrever a estrutura de um conjunto de dados [fijk] aquele que contém o menor número possível de parâmetros e apresenta um bom ajuste. A estatística $X^2 = 2 \sum_i \sum_j \sum_k (f_{ijk} \times \ln \frac{f_{ijk}}{E_{ijk}})$, tem distribuição assintótica X^2 com g graus de liberdade [2], onde g é o número de parâmetros eliminados; o ajuste é bom quando χ^2 for menor do que χ^2_g (crítico) para um nível de significância desejado.

A decisão sobre o modelo final adequado pode ser tomada seguindo um processo de eliminação de parâmetros um a um, a partir do modelo mais completo com todos os parâmetros, chamado modelo saturado. A cada passo é calculada a estatística X^2; em dois passos imediatamente sucessivos, onde no posterior um parâmetro foi retirado do modelo, calculam-se as estatísticas X^2 com

g-1 e g graus de liberdade. A diferença entre elas tem distribuição assintótica X^2 com 1 grau de liberdade e dá indicação sobre a manutenção ou não do parâmetro em questão no modelo. Se $\chi^2_g - \chi^2_1 = \chi^2_{g-1}$ for maior do que o χ^2_1 para um nível de significância desejado, o parâmetro é retido. A seqüência de testes se encerra quando todos os parâmetros remanescentes têm indicações para não serem retirados. Se os parâmetros B_{12}, B_{13}, B_{23} e B_{123} puderem ser eliminados restando o modelo $1n\ F_{ijk} = B + B_1(i) + B_2(j) + B_3(k)$, tem-se a situação de completa independência entre as três variáveis.

EXEMPLO DE AJUSTE; EXAME DAS ASSOCIAÇÕES

A Tabela 2 apresenta dados sobre gestantes quanto ao tabagismo (variável 3), sobre escolaridade da gestante (variável 2) e o peso do seu recém-nascido (variável 1). Cada uma das variáveis tem duas categorias mutuamente exclusivas: — não fuma (1), fuma (2); escolaridade alta (1), baixa (2); baixo peso: não (1), sim (2).

TABELA 2
Recém-nascidos vivos segundo peso ao nascer, tabagismos e escolaridade da mãe. Trinta e uma maternidades do município de São Paulo, 1978. *Modelo O.*

Baixo Peso	Escolaridade	Tabagismo			
		Não		Sim	
		Alta	Baixa	Alta	Baixa
Não		2.710	4.911	1.256	2.009
Sim		83	207	65	165

Fonte: Benicío, M.H.D.'A.[1]

O modelo log-linear completo, saturado, é o modelo 0. Modelo 0: $1\ n\ F_{ijk} = B + B_1(i) + B_2(j) + B_3(k) + B_{12}(ij) + B_{13}(ik) + B_{23}(jk) + B_{123}(ijk)$; o χ^2 para este modelo não é definido, pois o número de parâmetros é igual ao número de freqüências observadas. Há interesse em verificar qual dos parâmetros B_{123}, B_{12}, B_{13}, B_{23} deve permanecer. O primeiro passo é ajustar um modelo em que B_{123} (ijk) é eliminado; é o modelo 123: $1n\ F_{ijk} = B + B_1(i) + B_2(j) + B_3(k) + B_{12}(ij) + B_{13}(ij) + B_{23}(jk)$.

Calculado o χ^2 para o ajuste deste modelo, toma-se a decisão sobre a eliminação ou não de B_{123} (ijk). A Tabela 3 mostra as freqüências esperadas para este modelo; $\chi^2_1 = 0{,}509$ é um indicador de um bom ajuste.

TABELA 3
Freqüências esperadas sob o modelo 123.

Baixo Peso	Escolaridade	Tabagismo			
		Não		Sim	
		Alta	Baixa	Alta	Baixa
Não		2.713,6	4.907,4	1.252,4	2.012,6
Sim		79,4	210,6	68,6	161,4

$b_{12}^{(11)} = 0{,}095527704$
$b_{13}^{(11)} = 0{,}156531329$
$b_{23}^{(11)} = -0{,}029746431$

O modelo seguinte a ser ajustado é o modelo 123, 23, aquele em que foram retirados os parâmetros B_{123} e B_{23}, permitindo verificar se o parâmetro B_{23} deve ou não ser retirado. A Tabela 4 mostra as freqüências esperadas para este modelo; $\chi^2_2 = 8{,}328$ indica que o ajuste não é bom, ou seja, B_{23} deve permanecer. O teste de b_{23} pode ser feito mediante 8,328 − 0,509 = 7,819 que tem distribuição aproximada χ^2 com 1 grau de liberdade; o valor observado sugere que B_{23} é diferente de zero e, portanto, deve ser mantido.

TABELA 4

Freqüências esperadas sob o modelo 123, 23.

Baixo Peso	Escolaridade	Tabagismo			
		Não		Sim	
		Alta	Baixa	Alta	Baixa
Não		2.776,5	4.844,5	1.189,5	2.075,5
Sim		82,5	207,5	65,5	164,5

A retirada ou não do parâmetro B_{12} é decidida a partir do modelo 123, 12. A Tabela 5 mostra as freqüências esperadas; $\chi^2_2 = 15{,}956$ indica que o ajuste não é bom, ou seja, B_{12} deve ser mantido. O teste de b_{12} mediante a diferença de χ^2 tem o seguinte resultado: 15.956 − 0,509 = 15,447.

TABELA 5

Freqüências esperadas sob o modelo 123, 12.

Baixo Peso	Escolaridade	Tabagismo			
		Não		Sim	
		Alta	Baixa	Alta	Baixa
Não		2.690,6	4.930,4	1.234,1	2.030,9
Sim		102,4	187,6	86,9	143,1

Finalmente procura-se ajustar o modelo 123, 13 para verificar se o parâmetro B_{13} pode ser retirado. A Tabela 6 mostra as freqüências esperadas sob este modelo; $\chi^2 = 46{,}394$ indica que o parâmetro deve permanecer. O teste de b_{13} é 46,394 − 0,509 = 45,885.

TABELA 6

Freqüências esperadas sob o modelo 123, 13.

Baixo Peso	Escolaridade	Tabagismo			
		Não		Sim	
		Alta	Baixa	Alta	Baixa
Não		2.692,5	4.856,9	1.273,5	2.063,1
Sim		100,5	261,1	47,5	110,9

Portanto, o modelo final que permite um bom ajuste é o modelo 123. As estimativas b_{12}, b_{13} e b_{23} são feitas usando os E_{ijk} da Tabela 3 e são apresentadas no seu rodapé.

Tem-se as seguintes interpretações: 1) Quer para mães fumantes como para não-fumantes, há associação positiva entre baixa escolaridade da mãe e baixo peso ao nascer do filho — parâmetro B_{12} (11). 2) Qualquer que seja a escolaridade da mãe, há associação positiva entre a mãe fumar e baixo peso ao nascer — parâmetro B_{13} (11). 3) Há associação negativa entre escolaridade alta da mãe e ela não fumar, ou associação positiva entre escolaridade alta e fumar — parâmetro B_{23} (11). 4) Não há interação simultânea das três variáveis — parâmetro B_{123} (111).

RISCO RELATIVO; RAZÃO DOS PRODUTOS CRUZADOS

Seja a variável 1 considerada "resposta" e a distribuição das freqüências nas duas categorias desta variável resposta dependente das categorias das outras variáveis chamadas "fatores". A categoria 2 da variável 1 (nascimento de uma criança com baixo peso) pode ser considerada como representando um evento desfavorável; assim a relação $E_{2jk} \div (E_{1jk} + E_{2jk})$, estimadas por $f_{2jk} \div (f_{1jk} + f_{2jk})$, mede o "risco" de uma mãe com a combinação de características jk quanto às variáveis 2 e 3 vir a ter um evento desfavorável, qual seja, ter um recém-nascido de baixo peso.

É possível, e muitas vezes desejável, comparar riscos associados a diferentes combinações de categorias dos fatores. Por exemplo, no caso específico que está sendo apresentado, uma comparação seria entre riscos de baixo peso de recém-nascidos de mães que têm baixa escolaridade com mães que têm alta escolaridade, entre as não-fumantes. Usando os dados da Tabela 2, tem-se

$$[\, f_{221} \div (f_{121} + f_{221})\,] \div [f_{211} \div (f_{111} + f_{222})] = 1{,}36$$

O valor 1,36 é o *risco relativo* (estimado) e diz que o risco de uma mulher não-fumante de baixa escolaridade ter um recém-nascido de baixo peso é 1,36 maior do que o risco de uma mulher não-fumante de alta escolaridade. Uma boa aproximação de risco relativo é a *razão dos produtos cruzados* = *RPC*, onde

$$RPC = (E_{2jk} \times E_{1j'k}) \div (E_{1jk} \times E_{2j'k})$$

Usando logarítimo:

$$\ln \frac{E_{2jk}}{E_{1jk}} \text{ é o } \textit{logito} \text{ de } E_{2jk}$$

$$\ln \frac{E_{2j'k}}{E_{1j'k}} \text{ é o } \textit{logito} \text{ de } E_{2j'k}$$

Para a situação da tabela $2 \times 2 \times 2$ na configuração aqui apresentada, pode-se mostrar que, para as variáveis resposta (1) e fator (2), $\ln RPC(12) = 4 \times B_{12}$ (11), e para as variáveis resposta (1) e fator (3), $\ln RPC(13) = 4 \times B_{13}$ (11), desde que $B_{123} = 0$.

Vê-se que RPC até o momento foi avaliado relacionando o risco de categoria, "mais favorável" em relação à "menos favorável", da variável 2, dentro de cada uma das categorias da variável 3, assim como relacionando o risco da categoria "mais favorável" em relação à "menos favorável", da variável 3, dentro de cada uma das categorias da variável 2.

Pode-se também avaliar qual o risco relativo quando o indivíduo pertence simultaneamente às categorias "menos favoráveis" das variáveis 2 e 3, em comparação com indivíduo que pertence simultaneamente às respectivas categorias "mais favoráveis". No exemplo, é o risco relativo de ter recém-nascido de baixo peso entre mães de baixa escolaridade

que fumam e mães de alta escolaridade que não fumam. Em geral, existe este interesse, de comparar riscos de combinação de fatores desfavoráveis em relação a uma *categoria basal*, que é aquele em que as categorias dos fatores são as mais favoráveis. Para esta situação, tem-se

$$1\,n\,RPC = 4[(B_{12}(11) + B_{13}(11)]$$

MODELO LOGITO; OBTENÇÃO DA RAZÃO DOS PRODUTOS CRUZADOS

É possível ajustar-se um modelo logito a um conjunto de dados [fijk]. Existe equivalência de resultados entre modelo log-linear e modelo logito e igualdade de resultados quanto à obtenção de estimativas de razão de produtos cruzados, quando o modelo log-linear inclui todos os B's correspondentes a efeitos principais e aqueles correspondentes a todas as combinações possíveis de fatores, mais os B que contenham combinações da variável resposta com variável fator estatisticamente significante.

No caso de três variáveis com duas categorias cada, os modelos log-lineares de interesse que seriam equivalentes a modelos logitos são os modelos

$$1n\,F_{ijk} = B + B_1(i) + B_2(j) + B_3(k) + B_{23}(jk) + B_{12}(ij)$$

$$1n\,F_{ijk} = B + B_1(i) + B_2(j) + B_3(k) + B_{23}(jk) + B_{13}(ik)$$

$$1n\,F_{ijk} = B + B_1(i) + B_2(j) + B_3(k) + B_{23}((jk) + B_{12}(ij) + B_{13}(ik)$$

Ajustado um modelo log-linear, as razões dos produtos cruzados (estimativa dos riscos relativos) podem ser obtidas diretamente dos E_{ikj}. Assim, usando a Tabela 3, tem-se:

Risco relativo entre escolaridade alta e baixa $= (210{,}6 \times 2.713{,}6) \div (79{,}4 \times 4.907{,}4) = (161{,}4 \times 1.252{,}4) \div (2.012{,}6 \times 68{,}6) = e^{4b_{12}(11)} = 1{,}47$

Risco relativo entre fumantes e não-fumantes $= (2.713{,}6 \times 68{,}6) \div (79{,}4 \times 1.252{,}4) = (4.907{,}4 \times 161{,}4) \div (210{,}6 \times 2.012{,}6) = e^{4b_{13}(11)} = 1{,}87$

Risco relativo entre escolaridade baixa + fumantes e escolaridade alta + não-fumantes $= (2.713{,}6 \times 161{,}4) \div (2.012{,}6 \times 79{,}4) = e^{4[b_{12}(11)} + b_{13}(11)] = 2{,}74$

onde escolaridade alta + não-fumantes é a categoria basal.

Uma apresentação de resultados que pode facilitar a visão geral de relações é sob a forma da Tabela 7, onde se colocam as possíveis combinações de categorias das variáveis, a categoria basal e os respectivos riscos relativos. É subentendido que o risco relativo de combinações de categorias de variáveis é calculado em relação à categoria basal e que nas categorias que aparecem individualmente o risco é calculado em relação à categoria complementar. Costuma-se chamar tais categorias de "fatores de risco", com exceção da basal. A apresentação exemplificada na Tabela 7 é apropriada quando não há interação entre as três variáveis. Se houvesse interação ($B_{123} \times 0$), os riscos relativos de cada fator de risco seriam diferentes para cada categoria da outra variável.

TABELA 7
Riscos relativos segundo as variáveis

Categoria ("fator de risco")	Risco relativo
Basal (mães não-fumantes e de alta escolaridade)	1
Mães fumantes	1,87
Mães de baixa escolaridade	1,47
Mães fumantes e de baixa escolaridade	2,47

AGRADECIMENTO

A um dos relatores pelas valiosas sugestões.

SOUZA, J.M.P. de & BENÍCIO, M.H.D'A. Análise multivariada; um exemplo usando modelo log-linear. *Rev. Saúde públ.*, S. Paulo, **19**:263-9, 1985.

SOUZA, J.M.P. de & BENÍCIO, M.H.D'A. [Multivariate analysis — an example of the use of a log-linear model]. **Rev. Saúde públ.**, S. Paulo, 19:263-9, 1985.

ABSTRACT: A multivariate analysis of categorical data using a log-linear model for a 2 x 2 x 2 contingency table is presented.

UNITERMS: Multivariate analysis.

REFERÊNCIAS BIBLIOGRÁFICAS

1. BENÍCIO, M.H.D'A. Fatores de risco de baixo peso ao nascer em recém-nascidos vivos: município de São Paulo, 1978. São Paulo, 1983. [Tese de Doutoramento — Faculdade de Medicina da USP].

2. EVERITT, B.S. *The analysis of contingency tables.* New York, John Wiley & Sons, 1977.

3. BISHOP, Y.M.M.; FIENBERG, S.E. & HOLLAND, P.W. *Discrete multivariate analysis: theory and practice.* Cambridge, Mass., M.I.T. Press, 1975.

4. UPTON, G.J.G. *The analysis of cross: tabulated data.* New York, John Wiley & Sons, 1978.

5. WORCESTER, J. The relative odds in the 2^3 contingency table. *Amer. J. Epidem.*, **93**: 145-9, 1971.

6. VITALIANO, P.P. The use of logistic regression for modeling risk factors: with applications to non-melanoma skin cancer. *Amer. J. Epidem.*, **108**: 402-14, 1978.

Recebido para publicação em 27/12/1984

Aprovado para publicação em 21/03/1985