Utilização de estratificação e modelo de regressão logística na análise de dados de estudos caso-controle

Using of stratitification and the logistic regression model in the analysis of data of case-control studies

Suely Godoy Agostinho Gimeno, José Maria Pacheco de Souza

Departamento de Epidemiologia da Faculdade de Saúde Pública-Universidade de São Paulo - São Paulo, SP - Brasil

Exemplifica-se a aplicação de análise multivariada, por estratificação e com regressão logística, utilizando dados de um estudo caso-controle sobre câncer de esôfago. Oitenta e cinco casos e 292 controles foram classificados segundo sexo, idade e os hábitos de beber e de fumar. As estimativas por ponto dos odds *ratios* foram semelhantes, sendo as duas técnicas consideradas complementares.

Análise multivariada. Regressão logística. Estudos de casos e controles.

Em Epidemiologia, a regressão logística tem como objetivo descrever a relação entre um resultado (variável dependente ou resposta) e um conjunto simultâneo de variáveis explicativas (preditoras ou independentes), mediante um modelo que tenha bom ajuste, que seja biologicamente plausível e obedeça ao princípio da parcimônia[5]. Na análise estratificada tem-se o mesmo propósito, mas as relações são efetuadas uma a uma, isto é, somente é possível obter a estimativa do risco para um único fator de cada vez, controlando-se o conjunto das demais variáveis.

Podem ser citadas como vantagens da análise estratificada sua relativa simplicidade de execução, a facilidade de entendimento e maior proximidade que propicia entre o pesquisador e os dados. Entretanto, ao se aplicar essa técnica, o grau de dificuldade aumenta na medida em que aumenta o número de variáveis que tiverem que ser consideradas como controle; os testes de homogeneidade entre os estratos, para se verificar a existência de interação entre as variáveis, são feitos em etapa à parte; variáveis quantitativas não podem ser usadas em sua escala original; o nível global de significância é difícil de ser controlado[3,7].

A análise logística controla grande número de variáveis simultaneamente, permitindo que os dados sejam utilizados mais eficientemente; o teste de homogeneidade pode ser feito em conjunto, bastando introduzir no modelo o termo produto entre os fatores. Uma possível desvantagem é a eventual barreira que introduz entre o pesquisador e os dados; é praticamente obrigatório o uso de pacotes estatísticos e microcomputador[4, 5].

O presente trababalho exemplifica a utilização das análises estratificada e logística na análise de dados de estudos tipo caso-controle.

Material e Método

Foram utilizados dados de um estudo caso-controle sobre câncer de esôfago[6]. Oitenta e cinco casos de câncer de esôfago foram comparados com 292 controles hospitalares, classificados segundo segundo sexo, idade e os hábitos de beber e de fumar. O hábito de beber foi considerado fator de risco de principal interesse.

Foi verificada a existência de associação entre o câncer de esôfago e cada uma das variáveis, em uma primeira etapa (análise bruta), mediante a obtenção das estimativas dos *odds ratios* (OR), por ponto e por intervalo, além do valor da estatística quiquadrado de Mantel-Haenszel (χ^2_{MH})[1,2,9,10]. Nas etapas seguintes foram selecionadas as variáveis cujo

***Separatas/Reprints**: José Maria Pacheco de Souza - Av. Dr. Arnaldo, 715 - 01246-904 - São Paulo, SP - Brasil - Fax: (011) 282. 2920*
Edição subvencionada pela FAPESP. Processo 95/2290-6.
Recebido em 23.8.1994. Reapresentado em 22.5.1995. Aprovado em 12.6.1995.

valor do nível descritivo de significância do teste fosse menor do que 0,20[5].

A análise estratificada foi utilizada na obtenção da estimativa do *odds ratio* para o hábito de beber, controlando-se o efeito dos demais fatores previamente selecionados. Fez-se, quando possível, o teste de homogeneidade dos estratos, a fim de investigar a existência de interação entre as variáveis[2,9,10]. Este procedimento foi repetido utilizando-se o modelo de regressão logística (não condicional em virtude de não haver emparelhamento); a presença de interação entre as variáveis foi verificada mediante a introdução dos termos-produtos correspondentes. Os resultados das duas técnicas foram comparados.

Nos apêndices encontram-se as fórmulas utilizadas para a obtenção das estimativas dos *odds ratios*, assim como a descrição dos testes estatísticos. Foram utilizados os pacotes estatísticos para microcomputador Epi Info[3] e MULTLR[8]. Os intervalos de confiança para os estratos foram feitos segundo a técnica de Woolf[2].

Resultados

A Tabela 1 apresenta a distribuição completa dos casos e controles segundo sexo, idade e os hábitos de beber e de fumar. Na Tabela 2 encontram-se os resultados da análise bruta; a variável idade apresentou valor p > 0,20 e, dessa forma, não foi considerada nas etapas seguintes da análise. Nas Tabelas 3, 4, 5, 6 e 7 são apresentados os resultados das análises estratificada e logística, com uma e duas variáveis como controle.

Tabela 2- Resultados da análise bruta.

Variável	*Odds ratio*	χ^2_{MH}	p
Hábito de beber	6,91	35,88	0,00
Hábito de fumar	8,93	33,09	0,00
Sexo	3,41	17,83	0,00
Idade	0,80	0,82	0,37

Na Tabela 3 há sugestão de interação entre os hábitos de beber e de fumar e, também, parece haver ação protetora da bebida sobre a doença; como, sabidamente, o hábito de beber é importante fator de risco para o câncer de esôfago, torna-se indispensável a visualização dos dados no sentido de explicar o paradoxo. O pequeno número de casos entre os não-fumantes é responsável pela distorção observada; bastaria que a relação bebe: não bebe fosse 2:4 e o *odds ratio* seria 1,34. Na Tabela 4 os resultados são os esperados. Na análise estratificada, ao se considerar duas variáveis para a estratificação (Tabela 6), não foi possível fazer o teste de homogeneidade, devido a freqüência zero.

Nas Figuras 1, 2 e 3 estão apresentadas as estimativas, por ponto e por intervalo, dos *odds ratios* obtidos com a análise estratificada e com a regressão logística não condicional.

Comentários

Observou-se consistência entre os resultados obtidos com a aplicação das análises estratificada

Tabela 1- Casos e controles segundo sexo, idade e hábitos de beber e fumar. São Paulo, 1981.

Sexo	Idade	Hábito de beber	Hábito de fumar	Caso	Controle	Total
Feminino	≤ 57 anos	Não	Não	3	30	33
			Sim	-	15	15
		Sim	Não	-	14	14
			Sim	5	13	18
	> 57 anos	Não	Não	2	41	43
			Sim	3	8	11
		Sim	Não	-	6	6
			Sim	3	2	5
Masculino	≤ 57 anos	Não	Não	-	9	9
			Sim	2	12	14
		Sim	Não	1	8	9
			Sim	40	58	98
	> 57 anos	Não	Não	-	6	6
			Sim	-	19	19
		Sim	Não	-	4	4
			Sim	26	47	73
Total				85	292	377

Tabela 3 - Análise estratificada para hábito de beber, controlando-se hábito de fumar.

Estrato		Caso	Controle	Total	*Odds ratio*	Intervalo com 95% de confiança
Não fumante:	Bebe	1	32	33	0,54	0,06 - 4,78
	Não bebe	5	86	91		
Fumante:	Bebe	74	120	194	6,66	2,55 - 17,41
	Não bebe	5	54	59		
Total		85	292	377	4,50*	2,11 - 9,61

* Estimativa do *odds ratio* ponderado de Mantel-Haenszel (OR_{MH})
Teste de homogeneidade entre os estratos: χ^2_{1gl}= 4,27 (p = 0,04)

Tabela 4 - Análise estratificada para hábito de beber, controlando-se sexo.

Estrato		Caso	Controle	Total	*Odds ratio*	Intervalo com 95% de confiança
Feminino:	Bebe	8	35	43	2,69	0,94 - 7,71
	Não bebe	8	94	102		
Masculino:	Bebe	67	117	184	13,17	3,10 - 55,99
	Não bebe	2	46	48		
Total		85	292	377	6,28*	2,91 - 13,55

* Estimativa do *odds ratio* ponderado de Mantel-Haenszel (OR_{MH})
Teste de homogeneidade entre os estratos: χ^2_{1gl}= 3,03 (p = 0,08)

Tabela 5 - Análise com regressão logística não condicional para a variável hábito de beber, controlando-se (separadamente) hábito de fumar e sexo.

Modelo	Variável de controle	Hábito de beber *Odds ratio*	Intervalo com 95% de confiança
1	Hábito de fumar	4,04	1,94 - 8,46
2	Hábito de fumar	0,54	0,06 - 4,78
	Hábito de fumar e hábito de beber	12,39	2,55 - 17,40
3	Sexo	5,42	2,56 - 11,50
4	Sexo	2,69	0,94 - 7,71
	Sexo e hábito de beber	4,90	0,82 - 29,38

Tabela 6 - Análise estratificada para hábito de beber, controlando-se sexo e hábito de fumar.

Sexo	Hábito de fumar		Caso	Controle	Total	*Odds ratio*	Intervalo com 95% de confiança
Feminino	Não fumante:	Bebe	-	20	20	0	
		Não bebe	5	71	76		
	Fumante:	Bebe	8	15	23	4,09	0,93 - 17,92
		Não bebe	3	23	26		
Masculino	Não fumante:	Bebe	1	12	13	indefinido	
		Não bebe	-	15	15		
	Fumante:	Bebe	66	105	171	9,74	2,26 - 42,07
		Não bebe	2	31	33		
Total			85	292	377	4,79*	2,09 - 10,99

* Estimativa do *odds ratio* ponderado de Mantel-Haenszel (OR_{MH})
Teste de homogeneidade entre os estratos: não foi possível (caselas vazias).

e logística. Quando o número de variáveis a ser controlado simultaneamente é pequeno (uma ou duas), a análise estratificada é rápida, dispensa o uso de equipamentos eletrônicos e permite maior visibilização dos dados; resultados aparentemente paradoxais têm explicação quase que imediata. À medida em que o número de variáveis aumenta, a análise logística torna-se praticamente obrigatória, mesmo à custa de um possível distanciamento entre o pesquisador e os dados originais. Mas os dois tipos

Tabela 7 - Análise com regressão logística não condicional para as variáveis hábito de beber, hábito de fumar e sexo (simultaneamente).

Variáveis	*Odds ratio*	Intervalo com 95% de confiança
Hábito de beber (variável principal)	4,05	1,88 - 8,73
Hábito de fumar	5,01	1,92 - 13,12
Sexo	1,00	0,49 - 2,05

de abordagem não se excluem mutuamente[11]. O modelo logístico é mais flexível, com maior poder de exploração de variáveis. A existência de programas de microcomputador "amigáveis" e de uso livre, de abundância de cursos, de epidemiologistas com bom conhecimento de estatística, tornam cada vez mais conhecida e popular a análise logística. A análise estratificada será muitas vezes boa auxiliar para visibilização e compreensão das relações entre variáveis.

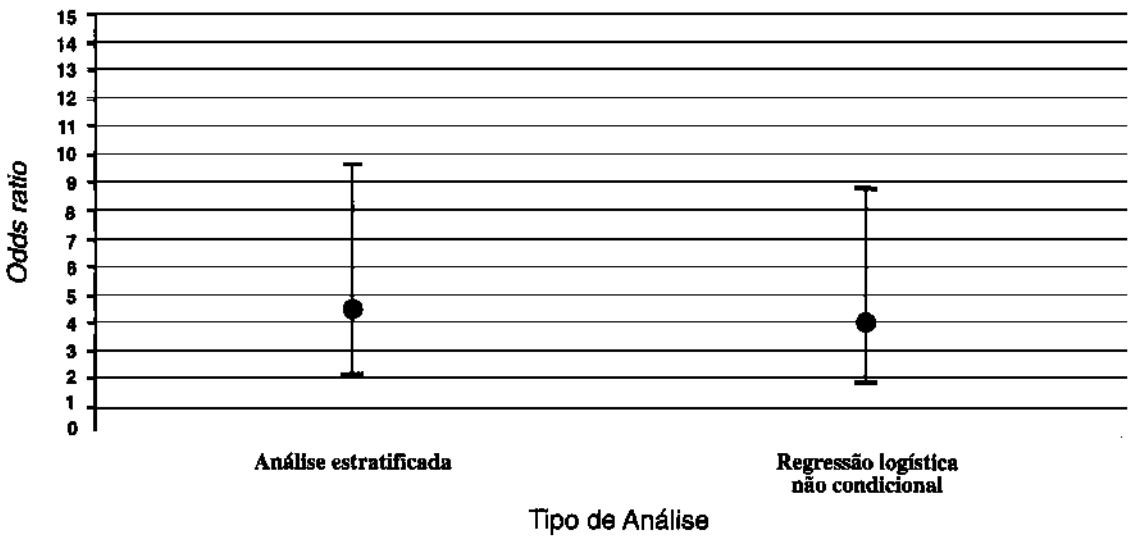

Figura 1- *Odds ratio* e intervalo com 95% de confiança para hábito de beber, controlando-se hábito de fumar, segundo tipo de análise.

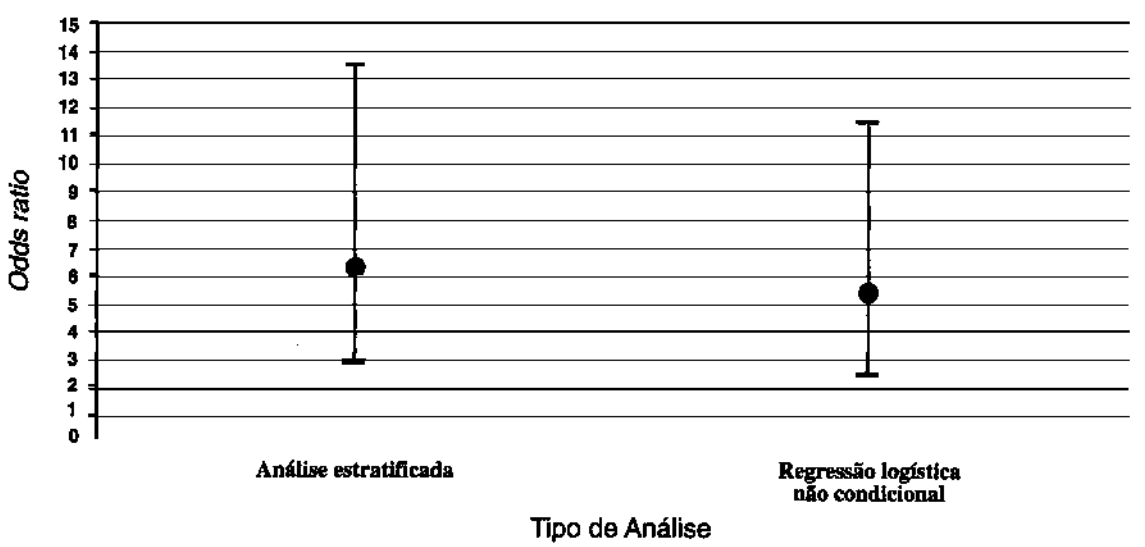

Figura 2 - *Odds ratio* e intervalo com 95% de confiança para hábito de beber, controlando-se sexo, segundo tipo de análise.

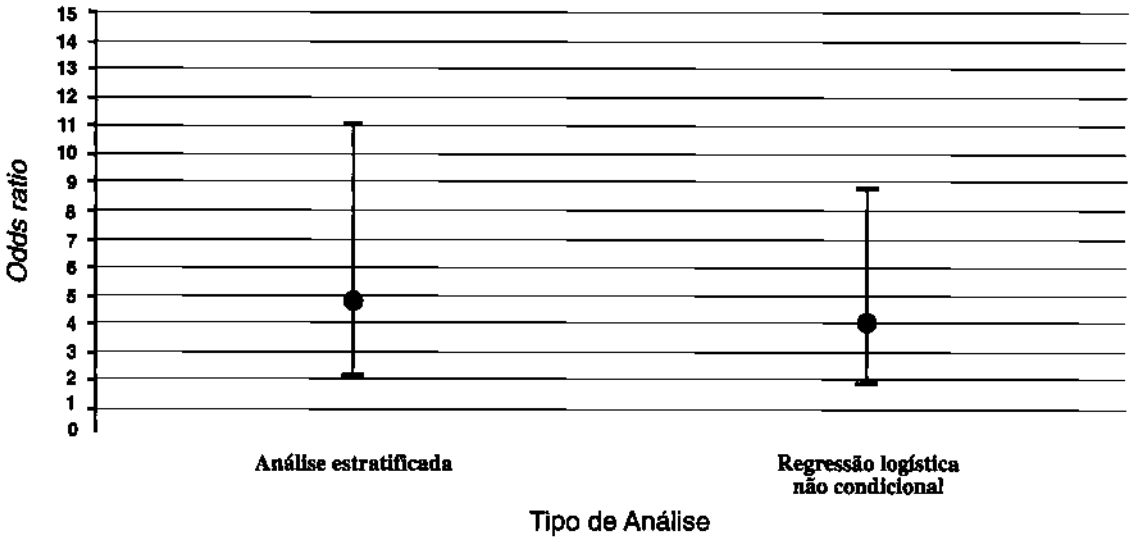

Figura 3 - *Odds ratio* e intervalo com 95% de confiança para hábito de beber, controlando-se sexo e hábito de fumar, segundo tipo de análise.

Agradecimentos

Aos relatores que apreciaram o manuscrito, pelas valiosas observações e sugestões.

Referências Bibliográficas

1. ARMITAGE, P. & BERRY, G. *Statistical methods in medical research*. 2nd. ed. Oxford, Blackwell Sci. Publ., 1987.
2. BRESLOW, N. E. & DAY, W. *Statistical methods in cancer research: the analysis of case-control studies*. Lyon, 1980, v. 1 (IARC Scient. Publ. n. 32).
3. DEAN, J.; DEAN, A.; BURTON, A.; DICKER, R. *Epi Info-computer programs for epidemiology*. Atlanta, Division of Surveilance and Epidemiologic Studies, Epidemiology Program Office, Center for Disease Control, 1990.
4. GREENLAND, S. Modelling and variables selection in epidemiologic analysis. *Am. J. Epidemiol.*, **124**: 869-76, 1986.
5. HOSMER, D. M. & LEMESHOW, S. *Applied logistic regression*. New York, John Wiley & Sons, 1989.
6. GIMENO, S. G. A. et al. Fatores de risco para o câncer de esôfago: estudo caso-controle em área metropolitana da região Sudeste do Brasil. *Rev. Saúde Pública*, **29**: 159-65, 1995.
7. MONCAU, J. E. C. Análise estratificada em estudos caso-controle. São Paulo, 1991. [Dissertação de Mestrado-Faculdade de Saúde Pública da USP].
8. MULTLR - A microcomputer program for multiple regression by condicional and uncondicional maximum likelihood methods. *Am.J.Epidemiol.*, **129**: 439-44, 1989.
9. ROTHMAN, K. J. *Modern epidemiology*. Boston, Little and Co., 1986.
10. SCHLESSELMAN, J.J. *Case-control studies - design, conduct, analysis*. New York, Oxford University Press, 1982.
11. VANDENBROUCKE, J. P. Should we abandon statistical modeling altoghether? *Am. J. Epidemiol.*, **126**: 10-3, 1987.

Abstract

Data of a case-control study of esophageal cancer were used as an example of the use of multivariate analysis with stratification and logistic regression. Eighty-five cases and 292 controls were classified according to sex, age and smoking and drinking habits. The point estimates of the odds ratios were similar, and the techniques were considered complementary.

Multivariate analysis. Logistic regression. Case-control studies.

Apêndice

I: Análise Bruta

Apresentação geral de uma Tabela formada por variáveis dicotômicas em estudo caso-controle

Fator de estudo	Condição		Total
	Caso	Controle	
Exposto	a	b	m_1
Não exposto	c	d	m_0
Total	n_1	n_0	T

Odds ratio = $\hat{OR} = \frac{ad}{bc}$

Intervalo de confiança para o *odds ratio*: $\exp\left[\ln(\hat{OR}) \mp z_{\alpha/2}\sqrt{V\left[\ln(\hat{OR})\right]}\right]$

Variância do $\ln(\hat{OR})$ = $V[\ln(\hat{OR})] = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$ (segundo Woolf)

Esperança de a = E(a) = $\frac{n_1 m_1}{T}$

Variância de a = V(a) = $\frac{m_1 n_0 m_0 n_1}{T^2 (T-1)}$

Teste de hipóteses: $\chi^2_{1gl} = \frac{[a - E(a)]^2}{V(a)}$

II: Análise Estratificada

Odds ratio de Mantel-Haenszel = $\hat{OR}_{MH} = \frac{\sum_{i=1}^{I} \frac{a_i d_i}{T_i}}{\sum_{i=1}^{I} \frac{b_i c_i}{T_i}}$

Intervalo de confiança para o *odds ratio* = $\exp\left[\ln(\hat{OR}) \mp z_{\alpha/2}\sqrt{V\left[\ln(\hat{OR})\right]}\right]$

$$V\left[\ln\left(\hat{OR}_{MH}\right)\right] = \frac{\sum P_i R_i}{2(\sum R_i)^2} + \frac{\sum P_i S_i + \sum Q_i R_i}{2(\sum R_i)(\sum S_i)} + \frac{\sum Q_i S_i}{2(\sum S_i)^2},$$

onde $R_i = \frac{a_i d_i}{T_i}$; $S_i = \frac{b_i c_i}{T_i}$; $P_i = \frac{a_i + d_i}{T_i}$; $Q_i = \frac{b_i + c_i}{T_i}$

Teste de homogeneidade entre I estratos:

$$\chi^2_{I-1gl} = \sum_{i=1}^{I} \frac{(a_i - A_i)^2}{V\left(A_i \mid \hat{OR}_{MH}\right)}, \text{ onde}$$

$$V\left(A_i \mid \hat{OR}_{MH}\right) = \frac{1}{\frac{1}{A_i} + \frac{1}{B_i} + \frac{1}{C_i} + \frac{1}{D_i}}$$

$$E(a_i \mid \text{homogeneidade}) = A_i = \frac{[(n_{1i} + m_{1i})\hat{OR}_{MH} + (n_{0i} - m_{1i})] - \sqrt{[(n_{1i} + m_{1i})\hat{OR}_{MH} + (n_{0i} - m_{1i})]^2 - 4n_{1i}m_{1i}\hat{OR}_{MH}(\hat{OR}_{MH} - 1)}}{2(\hat{OR}_{MH} - 1)}$$

$B_i = m_{1i} - A_i \qquad C_i = n_{1i} - A_i \qquad D_i = n_{0i} - B_i$

III: Análise Logística

Modelo sem Interação

Odds ratio = $\hat{OR} = \exp\left(\hat{\beta}\right)$

Modelo com interação

Odds ratio com interação:

estrato de não expostos à variável de controle: $\exp(\hat{\beta}_{principal})$

estrato de expostos a variável de controle: $\hat{OR}_{principal} \times \hat{OR}_{interação} = \exp(\hat{\beta}_{principal} + \hat{\beta}_{interação})$

Variância = $V(\hat{\beta}_{principal} + \hat{\beta}_{interação}) = V(\hat{\beta}_{principal}) + V(\hat{\beta}_{interação}) + 2\ \text{covariância}\ (\hat{\beta}_{principal}, \hat{\beta}_{interação})$

Intervalo de confiança: $\exp[(\hat{\beta}_{principal} + \hat{\beta}_{interação}) \pm z_{\alpha/2}\sqrt{\text{Variância}}]$