Avaliação da eficiência do método de Ward para comparação de modelos logísticos

Flávia Sílvia Corrêa Tomaz [1]- UFOP
Luiz Alexandre Peternelli - UFV
Sebastião Martins Filho - UFV

Introdução

Pesquisadores das mais diversas áreas deparam-se com a necessidade de comparar conjuntos de dados ou equações ajustadas para os vários tratamentos em estudo. Quando essas equações são consideradas idênticas, elas podem ser agrupadas em uma única expressão, o que simplifica a interpretação dos dados e a discussão dos resultados. Nesse sentido a literatura estatística apresenta alguns testes de identidade de modelos de regressão ou de alguns de seus coeficientes. Quanto à identidade de modelos não-lineares algumas referências podem ser citadas. Rao (1973) apresentou o teste da razão de verossimilhança com aproximação dada pela estatística qui-quadrado. Bates & Watts (1988) apresentou um teste assintótico baseado na razão de verossimilhança, com aproximação dada pela estatística F. Regazzi (2003) apresentou metodologia geral para a comparação de qualquer modelo não-linear baseado no teste da razão de verossimilhança. Regazzi e Silva (2004) utilizaram o teste da razão de verossimilhança com aproximação de qui-quadrado, no caso de delineamento inteiramente casualizado. No entanto, conforme discutido por Maia et al. (2009) em experimentos mais complexos algumas dessas propostas para identidade apresentam limitações.

Diante das limitações de algumas dessas técnicas e devido à simplicidade de se trabalhar com a análise de agrupamento tem-se percebido a utilização de métodos de agrupamento para a comparação de modelos. Portanto, esse trabalho tem como objetivo avaliar a eficiência do método de agrupamento de Ward na comparação de modelos logísticos.

[1] e-mail: flaviastomaz@yahoo.com.br

Metodologia

Análise de agrupamento

A análise de agrupamento, também conhecida como análise de conglomerado ou *cluster analysis* tem como objetivo dividir um conjunto de observações (elementos, indivíduos, tratamentos, genótipos, etc.) em grupos homogêneos ou compactos, segundo algum critério conveniente de similaridade. Assim, os elementos pertencentes a um mesmo grupo serão homogêneos (similares) entre si, com respeito a certas características medidas, enquanto que os pertencentes a grupos diferentes deverão ser heterogêneos entre si em relação às mesmas características (MINGOTI, 2005; SHARMA, 1996). O processo de agrupamento pode ser sintetizado em cinco etapas. A primeira é a escolha da medida de dissimilaridade, a seguinte é a escolha do método de agrupamento (hierárquico ou não-hierárquico). O terceiro passo é a escolha do tipo de agrupamento para o método escolhido seguido pela decisão sobre o número de grupos, e finalmente a interpretação do resultado do agrupamento (SHARMA, 1996; GNANADESIKAN, 1997).

Método de Ward

O método de Ward foi proposto por Ward (1963) e é também chamado de "Mínima Variância" (MINGOTI, 2005). Nesse método a formação dos grupos se dá pela maximização da homogeneidade dentro dos grupos. A soma de quadrados dentro dos grupos é usada como medida de homogeneidade. Isto é, o método de Ward tenta minimizar a soma de quadrados dentro do grupo. Os grupos formados em cada passo são resultantes de grupo solução com a menor soma de quadrados (SHARMA, 1996).

Simulação de dados

Para avaliar a eficiência do método de agrupamento de Ward na identidade de modelos logísticos foi realizado um estudo baseado em simulação de dados. Sendo que a eficiência corresponde ao número de vezes que o método de agrupamento separa corretamente os tratamentos dentro de cada grupo.

O modelo estatístico usado foi $y_i = \mu_i + \varepsilon_i = E(y_i) + \varepsilon_i$, onde a parte sistemática é dada pelo modelo logístico $E(Y) = \frac{1}{1+e^{\alpha(\beta - x)}}$. Nessa parametrização, α está relacionado com a taxa média de decrescimento da curva, e β corresponde ao ponto de

inflexão, ou seja, ponto correspondente a Y=50%. No modelo usado, considerou-se y_i como o valor gerado para uma observação dicotômica (HOSMER e LEMESHOW, 1989), com probabilidade π_i e $1-\pi_i$, para o resultado "sobrevivência", Y=1, ou "morte", Y=0. Assim, π_i representa a probabilidade de sobrevivência no i-ésimo tempo, e é calculada fazendo-se $\pi_i = E(y_i)$.

Para a definição dos cenários considerou-se o parâmetro $\alpha = -0{,}2456$, $\beta^*_1 = 10{,}28$ e β^*_2 definido de tal forma que $\beta^*_2 - \beta^*_1$ corresponda a porcentagens (p = 10, 20, 30, ..., 100%) de β^*_1, com diferentes tamanhos de amostra n = 10, 20, 30, 40, 50, 100. O parâmetro α foi considerado como um único valor devido à sua pequena influência no cálculo da matriz de dissimilaridade (TOMAZ, 2009).

Foram considerados 60 cenários que correspondem à combinação dos valores de β e *n*, juntamente com o valor único de α. Foram gerados oito tratamentos de forma que ao agrupá-los tivéssemos dois grupos compostos por quatro tratamentos cada. Assim β_1 corresponde ao parâmetro para o primeiro grupo gerado e β_2 para o segundo grupo gerado. Ou seja, os tratamentos 1, 2, 3 e 4 foram gerados utilizando $\alpha = -0{,}2456$ e $\beta_1 = 10{,}28$ e os tratamentos 5, 6, 7 e 8 a partir de $\alpha = -0{,}2456$ e β_2 variando conforme a diferença especificada acima.

Objetiva-se com este estudo verificar a relação entre a eficiência do agrupamento, tamanho da amostra e diferença entre os β de cada grupo de modelos. Para o estudo proposto foi desenvolvido uma rotina no software R® (R DEVELOPMENT CORE TEAM, 2008). Gerados os tratamentos, esses foram submetidos ao ajuste do modelo proposto $Y = \frac{1}{1+e^{\alpha(\beta - x)}} + \varepsilon$ e as estimativas dos parâmetros foram submetidas à análise de agrupamento, método de Ward. Como foi estabelecido o número de grupos a ser formado (2 grupos); o passo seguinte foi cortar o dendrograma gerado em dois grupos e avaliar se o agrupamento foi realizado corretamente. Isto é, se os tratamentos que foram gerados de um mesmo α e β pertenciam a um mesmo grupo. Foram realizadas 1000 simulações e registrado o número de vezes que o método de Ward conseguiu identificar corretamente os elementos dentro de cada grupo.

Resultados

A Tabela 1 mostra a proporção de vezes que os tratamentos foram agrupados corretamente. A Figura 1 apresenta a proporção de acertos no agrupamento dos tratamentos para amostras aleatórias de tamanhos 10, 20, 30, 40, 50, 100 em função da diferença entre os betas.

Tabela 1 - Proporção de acertos no agrupamento dos tratamentos para diferentes tamanhos de amostra e diferenças absolutas ($\beta_2^* - \beta_1^*$) e percentuais (**% de** β_1^*) relativas a β_2^* **e** β_1^*.

		Tamanho da Amostra					
$\beta_2^* - \beta_1^*$	% de β_1^*	10	20	30	40	50	100
1,028	10	0,086	0,192	0,327	0,444	0,551	0,839
2,056	20	0,445	0,775	0,91	0,959	0,986	0,998
3,084	30	0,821	0,978	0,999	1	1	1
4,112	40	0,966	0,997	1	1	1	1
5,14	50	0,996	1	1	1	1	1
6,188	60	1	1	1	1	1	1
7,196	70	1	1	1	1	1	1
8,224	80	1	1	1	1	1	1
9,252	90	1	1	1	1	1	1
10,28	100	1	1	1	1	1	1

Pela Tabela 1 e Figura 1 pode-se observar que:

1) Para pequenas diferenças entre β_2^* e β_1^*, isto é, $\beta_2^* - \beta_1^*$ = 10% de β_1^*, é necessário que o tamanho da amostra seja grande $(n \cong 100)$ para que o método de agrupamento de Ward classifique corretamente os modelos dentro de cada grupo;
2) Quando a diferença entre β_2^* e β_1^* é aproximadamente 4, ou seja, ($\beta_2^* - \beta_1^*$ = 40% de β_1^*), o método de agrupamento apresenta alta eficiência, uma vez que a proporção de acerto fica próximo a 1, mesmo quando o tamanho da amostra é pequeno;
3) Em geral, à medida que o tamanho da amostra aumenta a eficiência do agrupamento para a identidade de modelos também aumenta;
4) Quando $\beta_2^* - \beta_1^* > 5{,}14$, ou seja, $\beta_2^* - \beta_1^* \geq$ 50% de β_1^* a eficiência do método independe do tamanho da amostra.

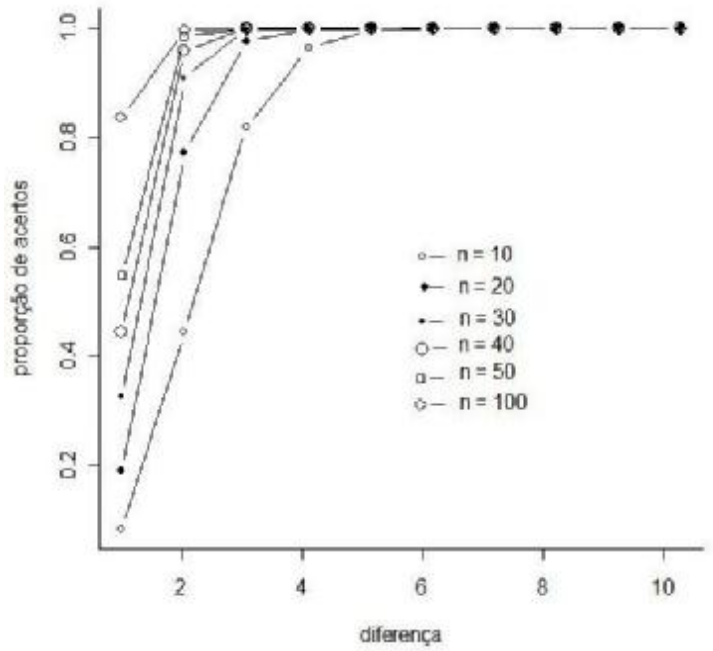

Figura 1 - Proporção de acertos em função do tamanho da amostra para uma dada diferença entre os valores dos betas.

Conclusões:

O método de agrupamento de Ward apresentou alto desempenho no agrupamento dos modelos logísticos quando as diferenças entre β_2^* e β_1^* foram superiores a 20% de β_1^* e o tamanho da amostra é igual ou superior a 30. O método proposto mostrou-se uma técnica em potencial para comparação de modelos logísticos.

Referências:

BATES, D. M.; WATTS, D.G. *Non linear analysis and its applications.* New York: John Wiley, 1988. 365p.

GNANADESIKAN, R. *Methods for statistical data analysis of multivariate observations.* 2. ed. New York, John Wiley an Sons, 1997.

HOSMER, D.W.; LEMESHOW, S. *Applied logistic Regression.* New York: John Wiley & Sons, 1989.

MAIA, E. *et al* .Método de comparação de modelos de regressão não-lineares em bananeiras. *Rev. Ciência Rural,* v.39, p. 1380-1386, 2009.

MATOS JÚNIOR, D.; GONZALES, A. F.; POMPEU JÚNIOR, J.; PARAZZI, C. Avaliação de curvas de maturação de laranjas por análise de agrupamento. *Pesq. agropec. bras*, v. 34, n. 12, p. 2203 – 2209. Dez. 1999.

MINGOTI, S. A.; *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*, Editora UFMG, 2005.

RAO, C. R. *Linear statistical inference and its applications*. New York: John Wiley, 1973. 522p.

REGAZZI, A. J. Teste para verificar a igualdade de parâmetros e a identidade de modelos de regressão não linear. *Rev. Ceres*, Viçosa, v.50, n. 266, p. 287, p. 9-26, 2003.

REGAZZI, A.J.; SILVA, C.H.O. Teste para verificar a igualdade de parâmetros e a identidade de modelos de regressão nãolinear. I. Dados no delineamento inteiramente casualizado. *Revista de Matemática e Estatística*, v.22, p.33-45, 2004.

SHARMA, S. *Applied multivariate techniques*. New York: John Wiley & Sons, 1996.

R DEVELOPMENT CORE TEAM. 2008. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Disponível em: <http://www.R-project.org>. Acesso em: 2008.

TOMAZ, F. S. C. *Análise de Agrupamento para avaliação de identidade de modelos não-lineares em análise de sobrevivência*. 2009. 70f. Dissertação (Mestrado em Estatística Aplicada e Biometria). Universidade Federal de Viçosa, Viçosa, 2009.

WARD, J. H. Hierarquical grouping to optimize an objective function. *Journal of the American Statistical Association*, v. 58, p. 236 – 244. Mar. 1963.