IMPACTO DE ERROS DE CLASSIFICAÇÃO EM DADOS MULTINOMIAIS

Octávio A. Torres[1], **Magda C. Pires**[1]

RESUMO

Introdução

A tarefa de classificar indivíduos segundo alguma característica está presente na maioria das áreas de conhecimento. Podemos, por exemplo, classificar as pessoas de acordo com a raça (branca, negra, parda etc), com o estado de saúde (doente ou não doente), com a preferência por candidato à eleição e muitos outros critérios. Nesse contexto, sabemos que as variáveis estudadas têm distribuição multinomial.

Entretanto, essa classificação pode estar sujeita a erros, ou seja, um indivíduo pode estar alocado a uma categoria que não corresponde ao seu estado verdadeiro. Consequentemente, a proporção estimada da categoria i será viciada.

Nesse contexto, o objetivo deste trabalho é ilustrar, através de simulações, o impacto dos erros de classificação na estimação das proporções de cada categoria das variáveis estudadas.

Metodologia

Variáveis aleatórias com distribuição Multinomial têm função de probabilidade dada por [5]:

$$f(y|n) = \frac{n!}{y_1!y_2!\ldots y_J!}\pi_1^{y_1}\pi_2^{y_2}\ldots\pi_J^{y_J} \qquad (1)$$

em que n é o tamanho da amostra, J o número de categorias, y_j é o número de observações na amostra pertencente à $j-$ésima categoria e π_j a probabilidade de ocorrência da $j-$ésima categoria.

[1]Universidade Federal de Minas Gerais - Departamento de Estatística,
octavio@ufmg.br

Seja λ_{ij} a probabilidade de que um indivíduo da categoria i seja classificado na categoria j. Para avaliar o impacto desses erros na estimação das proporções de interesse, foi desenvolvida uma rotina no software estatístico R [7], que consiste em gerar, a partir de uma amostra com classificação perfeita, $t = 1000$ amostras sujeitas a erros de classificação λ_{ij}. A partir das amostras geradas, calcula-se a proporção média e o desvio-padrão estimados para cada categoria.

Resultados

Neste trabalho realizamos simulações para variáveis com três e cinco categorias, tamanhos de amostra de $n = 20$ e $n = 200$ observações e probabilidades totais de erros de classificação de 0.05, 0.10, e 0.20 para cada categoria. No caso com três categorias, gerou-se uma amostra de distribuição Multinomial $(n, 0.1, 0.3, 0.6)$ e, no caso com cinco categorias, uma Multinomial $(n, 0.05, 0.15, 0.20, 0.25e0.35)$.

As Tabelas 1 e 2 apresentam, respectivamente, os resultados obtidos para três e cinco categorias.

Tabela 1: Estimativas obtidas para amostras de três categorias

Categoria (prob.)	Prob. total de erro	Número de categorias= 3 $n = 20$ Média	DP	$n = 200$ Média	DP
A(0.60)	0.05	0.5791	0.0435	0.5805	0.0138
	0.10	0.5596	0.0192	0.5596	0.0201
	0.20	0.5190	0.0252	0.5202	0.0257
B(0.10)	0.05	0.1167	0.0348	0.1175	0.0117
	0.10	0.1360	0.0162	0.1351	0.0163
	0.20	0.1698	0.0219	0.1693	0.0225
C(0.30)	0.05	0.3043	0.0411	0.3020	0.0125
	0.10	0.3045	0.0173	0.3053	0.0175
	0.20	0.3112	0.0231	0.3105	0.0234

Tabela 2: Estimativas obtidas para amostras de cinco categorias

Categoria (prob.)	Prob. total de erro	Número de categorias= 5 $n = 20$ Média	DP	$n = 200$ Média	DP
A(0.05)	0.05	0.0605	0.0266	0.0605	0.0087
	0.10	0.0732	0.0372	0.0712	0.0118
	0.20	0.0888	0.0538	0.0878	0.0163
B(0.15)	0.05	0.1531	0.0297	0.1534	0.0095
	0.10	0.1545	0.0408	0.1561	0.0134
	0.20	0.1591	0.0555	0.1629	0.0183
C(0.20)	0.05	0.1996	0.0329	0.1999	0.0101
	0.10	0.1978	0.0434	0.1998	0.0138
	0.20	0.1986	0.0585	0.2011	0.0187
D(0.25)	0.05	0.2466	0.0325	0.2467	0.0104
	0.10	0.2458	0.0463	0.2441	0.0140
	0.20	0.2362	0.0611	0.2373	0.0199
E(0.35)	0.05	0.3402	0.0329	0.3395	0.0085
	0.10	0.3288	0.0468	0.3288	0.0150
	0.20	0.3173	0.0640	0.3110	0.0204

Conclusões

Os resultados demonstram que as estimativas médias se tornam mais viciadas na medida em que a probabilidade total de erro aumenta e o tamanho da amostra diminui. O mesmo comportamento é observado para o desvio padrão.

Trabalhos Futuros

Demonstrado que a presença dos erros de classificação afeta a estimação dos parâmetros de interesse, pretende-se revisar os métodos clássicos [1, 3, 6] e bayesianos [4] já propostos na literatura para lidar com esse problema.

Em seguida, pretende-se avaliar o impacto dos erros de classificação da estimação dos parâmetros do modelo de Regressão Logística Nominal [2].

Referências

[1] Chen, T. T., A review of methods for misclassified categorical data in epidemiology. *Statistical in Medicine*,**8**, 1095-1106, 1989.

[2] Dobson, A. J., An introduction to generalized linear models. 2 ed. Boca Raton: Chapman & Hall, 2002.

[3] Fleiss, J. L.; Levin, B.; Paik, M. C., "Statistical Methods for Rates and Proportions", chapter: Misclassification: Effects, Control, and Adjustment. New York: Wiley, 1981.

[4] Paulino, C. D.; Turkman, M. A. A.; Murteira, B., "Estatística Bayesiana". Lisboa: Fundação Calouste Gulbenkian, 2003.

[5] Pérez, C. J.; Girón, F. J.; Martín, J.; Ruiz, M.; Rojano, C., Misclassified multinomial data: a Bayesian approach. *Rev. R. Acad. Cien. Serie A. Mat.*, **101(1)**, 71-80, 2007.

[6] Ross, S., "Probabilidade: Um curso moderno com aplicações", Bookman, Porto Alegre, 8ª edição, 2010.

[7] Walter, S. D.; Irwing, L. M., Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of Clinical Epidemiology*, **41(9)**, 923-937, 1988.

[8] Development Core Team, 2010.