Identificação de outliers por alocação das observações através do modelo Poisson aplicado ao número de casos de Aids diagnosticados no Brasil

Outliers indentification by allocation of observations through the Poisson model applied to the number of Aids cases diagnosed in Brazil

Tania Miranda Nepomucena[1 *]; Marcelo Angelo Cirillo[2]

Resumo

O método da análise robusta de variância proposto por Bertaccini e Varriale (2006) permite monitorar o efeito de *outliers* no processo de modelagem estatística. Para isso, utiliza-se a formação de subconjuntos, nos quais as unidades amostrais são alocadas, baseando-se em apenas uma inspeção dos dados. Com o propósito de estender este método para o modelo Poisson, o presente trabalho propõe monitorar o efeito de *outliers* no número de casos de AIDS diagnosticados no Brasil no período de 2003 a 2006. A metodologia proposta foi viável e é recomendável para dados de contagem, sendo, portanto, uma importante técnica de análise de dados para identificação de *outliers* em amostras, podendo ser aplicado a outros modelos generalizados com as devidas modificações na obtenção dos resíduos.
Palavras-chave: AIDS. Modelo Poisson. Outliers.

Abstract

The robust analysis variance method proposed by Bertaccini and Varriale (2006) allows the monitoring of the outliers effect in the statistic modeling process. As for that, the formation of subsets is used, where the sample units are allocated, based only on data inspection. With the purpose of extending this method to the Poisson model, this work intends to monitor the effect of outliers in the number of AIDS cases diagnosed in Brazil, from 2003 to 2006. The methodology proposed was viable and is recommended for counting data, being, therefore, an important data analysis technique used to identify outliers in samples. It can also be applied to other generalized models with appropriate changes in obtaining residuals.
Key-words: AIDS. Poisson Model. Outliers.

[1] Acadêmica do doutorado em Estatística e Experimentação Agropecuária da UFLA / Professora do Instituto Federal de Educação, Ciência e Tecnologia, Catu-BA. E-mail: tanianepomucena@yahoo.com.br
* Autor para correspondência
[2] Professor Adjunto do Departamento de Ciências Exatas da UFLA. E-mail: macufla@gmail.com

Introdução

No processo de modelagem estatística para realizar uma análise exploratória do conjunto de dados, possivelmente a amostra revelará a presença de *outliers*, que, em síntese, trata-se de elementos que não obedecem a um padrão do conjunto de dados ao qual eles pertencem.

Barnett e Lewis (1994) adotaram uma abordagem estatística para agrupar as causas de ocorrência de *outliers* durante a amostragem de dados:

· variedade inerente à população: os *outliers* são elementos que pertencem à população;

· erros de medição: ocorre na coleta dos dados. Pode ser causada por erros humanos, como a digitação de dados incorretos ou por erros de máquinas;

· erros de execução: ocorrem quando os dados são adquiridos por meio de amostragem de mais de uma população.

Os *outliers* requerem atenção especial, pois normalmente essas observações resultam de alguma violação das pressuposições necessárias para adequação ao modelo, produzindo conseqüentemente efeitos não confiáveis na eficiência dos estimadores (TUKEY, 1960).

Contextualizando a importância de estudar métodos que proporcionam um controle sobre a presença de *outliers*, podem ser citados vários estudos estatísticos sobre a disseminação da AIDS no Brasil. Neste sentido, encontraram-se estudos cujos resultados têm como principal objetivo o controle mais eficaz da epidemia no país.

Para um melhor entendimento do decurso da epidemia nas categorias compreendidas entre indivíduos com diferença de comportamento sexual, usuários de drogas, hemofílicos, receptores de sangue e crianças, Barbosa e Struchiner (1998), efetuaram um estudo de análise de sobrevida com aplicação do método Kaplan-Meyer e também do modelo de regressão Poisson, e estimaram o número de casos de AIDS diagnosticados no Brasil desde 1986 a junho de 1996, baseados na correção de atraso da notificação. Nessa mesma pesquisa, ambas as metodologias ainda sinalizaram particularidades relacionadas ao crescimento diferenciado da epidemia entre as categorias, além de diferença nas estimativas de probabilidade do atraso de notificação entre grupos.

O alcance de maior sobrevida dos portadores de AIDS também é um assunto presente entre várias pesquisas realizadas sobre a epidemia. Signorini et al. (2005) investigaram a influência dos fatores sociodemográficos, clínico-profiláticos e terapêuticos na vida de pacientes do hospital universitário do Rio de Janeiro, após o diagnóstico de AIDS, no período de 1995 a 2002. Baseados nas estimativas do efeito de algumas variáveis de caráter clínico e social consideradas no estudo, a modelagem Kaplan-Meyer e o modelo de Cox revelaram importantes informações que podem orientar em aspectos relacionados ao acompanhamento do paciente para o sucesso no livre acesso à terapia anti-retroviral do controle da AIDS.

Brito et al. (2006) realizaram uma análise estatística com modelos de regressão exponencial ajustados à série temporal, usando dados referentes ao número de casos de AIDS notificados em indivíduos com idade igual ou inferior a 13 anos, por transmissão vertical, no período de 1990 a abril de 2004. Assim, os autores concluíram que a partir de 1997, período que coincide com a inserção de medidas profiláticas preconizadas pelo Programa Nacional de DST e AIDS do Ministério da Saúde, houve uma redução progressiva, em todas as regiões do Brasil, do número de casos de AIDS previstos por transmissão vertical.

Com base nos trabalhos relatados anteriormente, salienta-se que diversas metodologias estatísticas poderão ser utilizadas para realizar estudos sobre

diagnóstico e prevenção da AIDS. Entretanto, todas elas devem levar em consideração um estudo preliminar do efeito das observações classificadas como *outliers*. Algumas metodologias são encontradas na literatura relacionada à estimação robusta (BARNETT, 1988; ATKINSON; RIANI, 2000). Porém, convém ressaltar que o processo de estimação robusta torna-se um pouco complexo, pois exige do pesquisador conhecimentos mais específicos sobre teoria de estimação. Dessa forma, tem-se a motivação para pesquisa de novas metodologias que sejam de melhor entendimento e possam ser implementadas computacionalmente com facilidade.

Com o intuito de monitorar o efeito das observações *outliers* sobre as estimativas de um modelo, Bertaccini e Varriale (2006) propuseram um método para detectar e investigar o efeito destas observações, considerando um modelo linear utilizado em conjunto com a técnica de análise de variância. A metodologia proposta por esses autores permitiu não só analisar seus efeitos na estimação de parâmetros, mas também verificar o desempenho em testes de significância relacionados aos parâmetros de interesse. Para isso, ajusta-se o modelo linear para um número relativamente pequeno de observações, meramente selecionadas por uma simples inspeção dos dados. Em seguida, sequencialmente, novas observações são introduzidas e novamente o modelo é reajustado. Dessa forma, procede-se com o reajuste do modelo para cada nova observação inserida, até que todos os dados sejam analisados.

Em virtude do que foi mencionado, este trabalho propõe estender o método apresentado por Bertaccini e Varriale (2006) para o modelo Poisson, com o propósito de identificar *outliers* através da alocação de observações na amostra referente ao número de casos de AIDS diagnosticados por estado brasileiro, no período de 2003 a 2006.

Material e métodos

Para a realização deste trabalho, consideraram-se os dados referentes ao número de casos de AIDS diagnosticados nos estados brasileiros, no período de 2003 a 2006, divulgados via internet, pelo Departamento de Informática do SUS - DATASUS, órgão da Secretaria Executiva do Ministério da Saúde.

Com o objetivo de identificar *outliers* verificados nos dados amostrais, estendeu-se para o modelo Poisson (1), o método proposto por Bertaccini e Varriale (2006) que, em resumo, trata-se da construção de uma análise robusta de variância em modelos lineares.

Para a formulação do modelo, considerou a observação como o número de casos de AIDS diagnosticado no i-ésimo ano (i=1,...,4) no j-ésimo estado (j = 1,..., 27). Desta forma, obteve-se o modelo Poisson por (1).

$$\begin{aligned} &E\left(y_{ij}\right)= e^{\eta_{ij}} = \mu_{ij} \text{ em que,} \\ &\eta_{ij} = \ln\left(\mu_{ij}\right)= s_{ij} \end{aligned} \qquad (1)$$

sendo η_{ij} o preditor linear, no qual considerou-se a média μ_{ij} e s_{ij} é a estimativa do parâmetro referente ao i-ésimo ano (i=1,...,4) na j-ésima repetição (j=1,...,27).

O monitoramento dos *outliers* teve como ponto de partida o ajuste do modelo Poisson (1) para um subconjunto inicial com algumas observações selecionadas. A seleção dessas observações foi conduzida pela estratégia denominada alocação não-proporcional. A descrição do algoritmo utilizado para a alocação não-proporcional é dada conforme os seguintes passos:

Etapa 1: Escolha do conjunto inicial S_m

Determinou-se um subconjunto S_m de unidades amostrais, selecionando em cada grupo i, a j-ésima observação y_{ij} que apresentou a menor diferença em módulo, de acordo com a seguinte condição:

$$\min \left| y_{ij} - med_i \right|, \; i = 1, \ldots, 4 \text{ e } j = 1, \ldots, 27. \quad (2)$$

sendo a mediana amostral do i-ésimo grupo.

Etapa 2: Adicionando observações ao subconjunto S_m

Ajustou-se um modelo (1) para o subconjunto S_m, em seguida calculou-se para as demais observações o quadrado do resíduo ordinário

$$e^2 = (y_{ij} - \hat{y}_{ij})^2 . \quad (3)$$

A observação que apresentou o menor resíduo (3) foi inserida ao subconjunto S_m. Na realização dessa etapa, obteve-se um novo subconjunto de observações, denominado por $S_{(m+1)}$, e o processo mais uma vez foi repetido. O procedimento finalizou quando todas as observações foram inseridas no modelo. É importante ressaltar que o índice é obtido da igualdade $m = \sum_{i=1}^{4} m_i$, dado que os $m_i's$ são o número de observações do grupo i no passo m.

Etapa 3: Estimativas de parâmetros no monitoramento dos *outliers*

Para cada subconjunto obtido em $S_{(m+1)}$ computaram-se as estimativas de máxima verossimilhança dos parâmetros, seus respectivos erros padrões e a deviance do modelo, além dos resíduos ordinários (3) obtidos por meio da realização do ajuste em cada subconjunto gerado. Com base nestas informações, foram construídos gráficos, os quais permitiram avaliar o efeito dos *outliers* em relação ao modelo proposto (1).

Resultados e discussão

Uma análise preliminar das observações referentes ao número de casos de AIDS diagnosticados nos estados brasileiros em cada ano foi realizada conforme ilustra o gráfico boxplot (Figura 1).

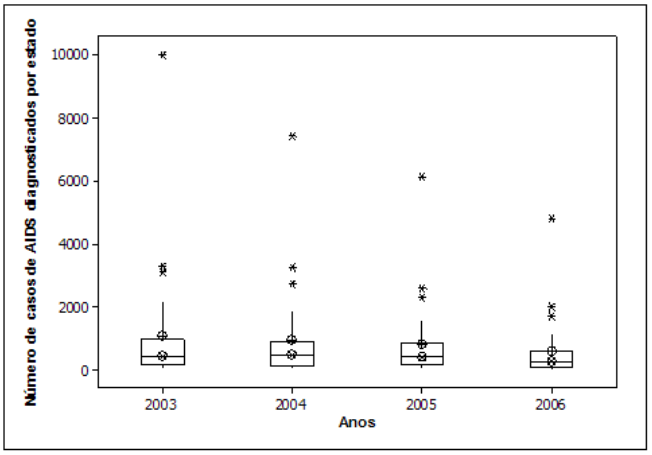

Figura 1. Boxplot dos números de casos de AIDS diagnosticados nos estados brasileiros no período de 2003 a 2006.

Analisando os resultados encontrados na Figura 1, nota-se que, em cada ano, houve a ocorrência de observações discrepantes. Essas observações são referenciadas na figura pelo símbolo "*". Importante ressaltar que essa análise conduz a aplicação dos procedimentos robustos, caracterizados nesse trabalho pela estratégia de alocação não-proporcional, considerando o modelo Poisson com a respectiva função de ligação logarítmica.

Aplicando a estratégia de alocação não-proporcional, os resultados ilustrados na Tabela 1 evidenciaram que o número de elementos

considerados em cada subconjunto amostral S_m, identificado pelos passos 20º até 67º, apresentaram comportamento que revela a presença de observações com magnitude bem diferenciada. Esse fato pode ser compreendido ao observar algumas repetições (estados), nas quais alguns valores diferiram substancialmente. É possível exemplificar esta situação, ao comparar o número de casos de AIDS do Acre com o número de casos do Distrito Federal no ano 2003, correspondendo a 35 e 527, respectivamente.

Tabela 1 - Subconjuntos das observações referentes ao número de casos de AIDS diagnosticados nos estados brasileiros, formados nos 20º a 67º passos da execução da estratégia não-proporcional para monitorar outliers usando o modelo Poisson.

Anos	20º passo	Anos	67° passo
2003	405, 467, 527, 427, 510	2003	35, 280, 35, 148, 68, 405, 67, 35, 467, 211, 163, 187, 139, 527, 289, 427, 510
2004	409, 460, 393, 511, 500	2004	42, 409, 44, 140, 95, 615, 77, 460, 203, 180, 123, 126, 393, 276, 511, 500
2005	425, 425, 498, 331	2005	48, 425, 57, 122, 52, 498, 49, 212, 606, 331, 191, 202, 148, 133, 346, 188, 431, 425
2006	128, 153, 184, 216, 232, 263, 106, 99	2006	22, 329, 48, 106, 51, 359, 46, 153, 376, 99, 128, 84, 69, 216, 184, 232, 263

A Figura 2 mostra como o tamanho dos subconjuntos aumenta a cada passo da alocação não-proporcional, à medida que observações são inseridas nos subgrupos amostrais. As oscilações mais acentuadas das linhas nos passos 20º e 67º indicam que possivelmente foi iniciada a inserção de outliers nos subconjuntos dos dados.

De acordo com os resultados identificados na Figura 1, mantendo o enfoque nos subconjuntos formados nos passos 20º até 67º (Tabela 1), realizou-se a análise do efeito dos outliers em relação ao erro padrão, ao desvio e às estimativas dos parâmetros do modelo Poisson (1).

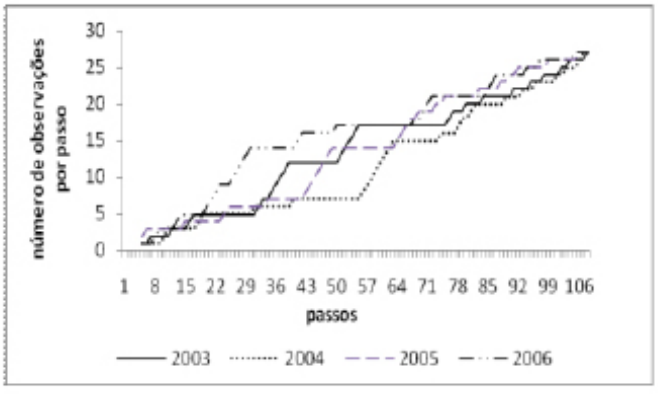

Figura 2. Número de observações inseridas em cada subconjunto amostral para os anos avaliados na estratégia de alocação não-proporcional

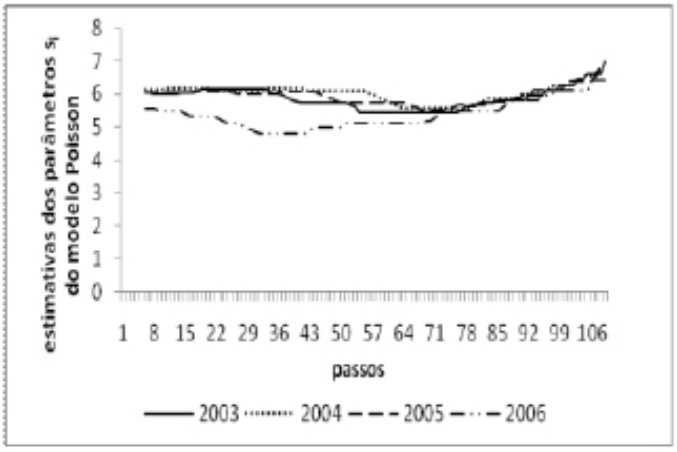

Figura 3. Estimativas dos parâmetros do modelo Poisson obtidas a cada passo da alocação não-proporcional de observações referentes ao número de casos de AIDS diagnosticados no Brasil no período de 2003 a 2006

Analisando o efeito dos outliers em relação às estimativas dos erros padrões, os resultados encontrados na Figura 4 permitiram verificar que as estimativas dos erros padrões referentes aos anos decresceram à medida que os subgrupos amostrais foram recebendo novas observações. O referido comportamento sinaliza a possibilidade de subestimação dos erros padrões, e conseqüentemente, uma avaliação incorreta das estimativas dos parâmetros dos modelos, conforme previsto pela literatura (DOBSON, 2002).

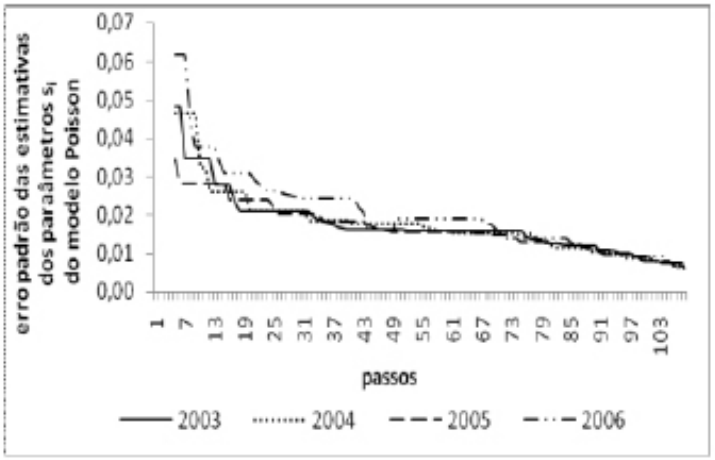

Figura 4. Estimativas dos erros padrões dos parâmetros do modelo Poisson obtidas a cada passo da alocação não-proporcional de observações referentes ao número de casos de AIDS diagnosticados no Brasil no período de 2003 a 2006.

A partir dos resultados da Figura 5, é mostrado que existem fortes evidências para que o modelo seja rejeitado, visto que o desvio foi determinado por altos valores em relação ao número de graus de liberdade, sinalizando a existência de variabilidade acentuada nos subgrupos amostrais.

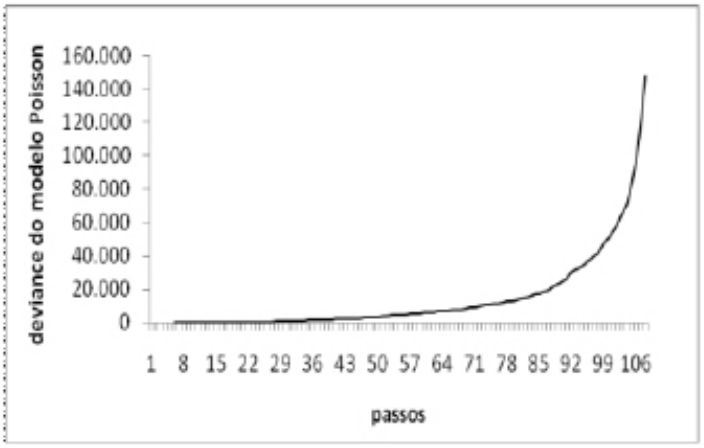

Figura 5. Inspeção dos valores da deviance do modelo Poisson utilizado a cada passo da alocação não-proporcional de observações referentes ao número de casos de AIDS diagnosticados no Brasil no período de 2003 a 2006.

Conclusão

A metodologia proposta neste estudo para investigação de outliers por meio do modelo Poisson em conjunto com aplicação da estratégia de alocação não-proporcional foi viável e é recomendável para dados de contagem, sendo, portanto, uma importante técnica quando se tem interesse em identificar outliers em dados amostrais.

Referências

ATKINSON, A. C.; RIANI, M. Robust diagnostic regression analysis. New York: Springer, 2000. p. 22-25.

BARBOSA, M. T. S.; STRUCHINER, C. J. Estimativas do número de casos de AIDS no Brasil. Revista Brasileira de Epidemiologia, São Paulo, v. 1, n. 3, p. 235-244, 1998.

BARNETT, V. Outlier and order statistics. Communications in Statistics: part A: theory and methods, New York, v. 17, n. 7, p. 2109–2118, 1988.

BARNETT, V.; LEWIS, T. Outliers in statistics. 3. ed. New York: J. Wiley, 1994. p. 3-18.

BERTACCINI, B.; VARRIALE, R. Robust analysis of variance: an approach based on the forward search. Computational Statistics & Data Analysis, Amsterdam, v. 51, n. 10, p. 5172-5183, 2006.

BRITO, A. M.; SOUSA, J. L.; LUNA, C. F.; DOURADO, I. Tendência da transmissão vertical de AIDS após terapia anti-retroviral no Brasil. Revista de Saúde Pública, São Paulo, v. 40, p. 18-22, 2006.

DOBSON, A. J. An introduction to generalized linear models. 2. ed. London: Chapman & Hall, 2002. p. 89-90.

SIGNORINI, D. J. H. P.; CODEÇO, C. T.; CARVALHO, M. S.; CAMPOS, D. P.; MONTEIRO, M. C. M.; ANDRADE, M. F. C.; PINTO, J. F. C.; SÁ, C. A. M. Effect of sociodemographic, clinical-prophylatic and therapeutic procedures on survival of AIDS patients assisted in a Brazilian outpatient clinic. Revista Brasileira de Epidemiologia, São Paulo, v. 8, n. 3, p. 253-261, 2005.

TUKEY, J.W. A survey of sampling from contaminated distribution: contributions to probability and statistics. California: University Stanford, 1960. p.448-485.

Recebido em 12 de maio de 2008 - Received on May 12, 2008.
Aceito em 17 de novembro de 2009 - Accepted on November 17, 2009.