Identification of Outlying Cells in Multi-way Tables

Jong Cheol Lee
Manager, UNIBOSS Inc.
23-5 Yoido-Dong, Youngdeungpo-Ku, Seoul, 150-110, Korea
jongskku@orgio.net

Chong Sun Hong
Professor, Department of Statistics Sungkyunkwan University
3-53 Myongryun-Dong, Chongro-Ku, Seoul, 110-745, Korea
cshong@skku.ac.kr

1. Estimation and Identification for Outlying Cells

When a particular model is tested for fitting contingency table, it is known that some outlying cells might cause significant lack of fit. Various criteria are suggested in order to identify these cells. One kind of these criteria is consisted of residuals, and another is of the difference of goodness-of-fit test statistics.

To detect more than one outlying cell, there are two identification methods, among others; one is the forwards-stepping method of Fuchs and Kenett (1980) which tests from the most extreme cell to the least extreme, and the other is the backwards-stepping method of Simonoff (1988) which tests from the least extreme cell to the most extreme.

In this paper, we propose a method which could identify several outlying cells in multi-way tables. This method uses an iterative proportional fitting (IPF) method to estimate suspected outlying counts. Since this iterative method uses minimal sufficient statistics of certain log-linear model at each iteration steps, one could estimate suspected outlying cell counts under any hierarchical log-linear models. With this method, more than one omitted cell counts could be estimated at once even in the case of multi-way tables.

By using the deleted residuals, the proposed identification method enables to detect multiple outlying cells for multi-way tables. Since this method is a modified version of the backwards-stepping method, both the masking and the swamping effects might be involved in this method. Nonetheless this method needs smaller number of iterations of calculation than that of the backwards-stepping. The proposed estimation and identification methods are explored with simulation studies.

2. Multiple Outlying Cells Identification

Simonoff (1988) propose the backwards-stepping method to identify outlying cells from the least extreme cell to the most extreme cell in $I \times J$ contingency tables. To reduce the swapping effects attributed to adjusted residuals of Haberman (1973), he use the deleted residuals in (2) instead of adjusted residuals to detect outlying cells. As the number of minimal sufficient statistics increases, the backwards-stepping method takes longer time to detect outlying cells because of many more number of calculations.

We propose an alternative method to identify multiple outlying cells at once. Moreover this method might have smaller number iterations than that of the backwards-stepping, and can be applied to general log-linear models for multi-way tables.

In this paper, some suspected outlying cells having extreme deleted residuals are put into the set S_q, and the set S_{q+1} is consisted with next considering outlying cells. By comparing S_q with S_{q+1}, outlying cells could be identified. That is, testing hypotheses could be obtained as

$$\begin{aligned} H_0 &: \textit{All elements in } S_{q+1} \textit{ are outlying cells.} \\ H_1 &: \textit{All elements in } S_q \textit{ are outlying cells.} \end{aligned} \tag{1}$$

The test statistic for the hypotheses (3) is

$$\Delta G_q^2 = G_{S_{q+1}}^2 - G_{S_q}^2, \tag{2}$$

which is the difference between two generalized likelihood ratio test statistics. One notes that this statistic (4) is identical with Cook's D-statistic to measure the influence of the corresponding cells (Christensen, 1990). Hence the hypotheses (3) could be transformed to

$$\begin{aligned} H_0 &: \textit{The cells which do not belong to } S_{q+1} \textit{ but } S_q \\ &\qquad \textit{are not influential cells.} \\ H_1 &: \textit{The cells which do not belong to } S_{q+1} \textit{ but } S_q \\ &\qquad \textit{are influential cells.} \end{aligned}$$

The test statistic ΔG_q^2 in (4) follows χ^2-distribution with some degrees of freedom which is the difference between the number of elements in two sets S_q and S_{q+1}. When ΔG_q^2 has smaller value, it means that there is no big difference between two goodness-of-fits. Hence elements (cells) in S_{q+1} might be considered as outlying cells. On the other hand, ΔG_q^2 having larger value means that elements in S_q might be regarded as outlying cells. And it could conclude that some elements which do not belong to S_{q+1} but S_q are not outlying cells. In this method, we select as many suspected outlying cells as possible at the initial step in order to reduce the masking effects. Thus, this method tests from the least extreme cell to the most extreme with similar arguments of the backwards-stepping method. Therefore, we knew that the number of initially suspected outlying cells in S_1 is larger than that of S_2 which in turn is larger than that of S_3 ....

REFERENCES

Barnett, V. and Lewis, T. (1994). *Outliers in statistical data. 3rd. John Wiley & Sons.*

Brown, M. L. (1974). Identification of the sources of significance in two-way contingency tables. *Applied Statistics.*, **23**, 405-413.

Christensen, R. (1990). *Log-linear models. New York : John Wiley & Sons.*

Fienberg, S. E. (1969) Preliminary Graphical analysis and quasi-independence for two-way contingency table. *Applied Statistics.*, **8**, 153-168.

Fuchs, C. and Kenett, R. (1980). A test for outlying cells in the multinomial distribution and two–way contingency tables. *Journal of the American Statistical Association.*, **75**, 395-398.

Goodman, L. A. (1968). The analysis of cross-classified data: independence, quasi-independence, and interaction in contingency tables with or without missing cells. *Journal of the American Statistical Association.*, **63**, 1091-1131.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics.*, **29**, 205-220.

Kotze, T. J. W. and Hawkins, D. M. (1984). The identification of outliers in two-way contingency tables using 2×2 subtables. *Applied Statistics.*, **33**, 215-223.

Mosteller, F. and Parunak, A. (1985). Identifying extreme cells in a sizable contingency table : probabilistic and exploratory approaches. *In Exploring Data Tables, Trends and Shapes. Wiley*, 189-224.

Simonoff, J. S. (1988). Detecting outlying cells in two-way contingency tables via backwards-stepping. *Technometrics.*, **30, 3**, 339-345 (with Correction, Vol. 32, No. 1, 115).

Upton, G. J. G. and Guillen, M. (1995). Perfect cells, direct models and contingency table outliers. *Communications in Statistics, Part A - Theory and Methods.*, **24**, 1843-1862.

RESUME

On propose une méthode d'identification pour découvrir plus d'un élément isolé dans une table d'éventualité de multi-dimension. Une méthode répétitive proportionnelle appropriée est appliquée pour obtenir des valeurs attendues de plusieurs éléments soupçonnés comme isolés. Dans la mesure oùla méthode proposée utilise une minimal statistique suffisant sous le modèle quasi log-ligne, le compte attendu des éléments isolés peut ẽtre évalué sous n'importe quel modèle quasi log-ligne hierarchique. Cette méthode est une extension de 'méthode de recul' chez Simonoff (1988) et demande une petite répétition pour identifier les éléments isolés.