

Decomposition of Independence Using the Logit Uniform Association Model and Equality of Concordance and Discordance for Two-Way Classifications

Kouji Tahata, Nobuko Miyamoto, Sadao Tomizawa

Department of Information Sciences, Tokyo University of Science, Chiba, Japan

Email: kouji_tahata@is.noda.tus.ac.jp, miyamoto@is.noda.tus.ac.jp, tomizawa@is.noda.tus.ac.jp

Received 4 September 2015; accepted 11 October 2015; published 16 October 2015

Copyright © 2015 by authors and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

For two-way contingency tables with ordered categories, the present paper gives a theorem that the independence model holds if and only if the logit uniform association model holds and equality of concordance and discordance for all pairs of adjacent rows and all dichotomous collapsing of the columns holds. Using the theorem, we analyze the cross-classification of duodenal ulcer patients according to operation and dumping severity.

Keywords

Concordance, Discordance, Independence, Logit Uniform Association Model

1. Introduction

Consider the $r \times c$ contingency tables with ordered categories, let X and Y denote the row and column variables, and let $P(X = i, Y = j) = p_{ij}$ (> 0) for $i = 1, \dots, r$ and $j = 1, \dots, c$. Goodman [1] considered the uniform association (U) model which was defined by

$$p_{ij} = \mu \alpha_i \beta_j \theta^{ij} \quad (i = 1, \dots, r; j = 1, \dots, c).$$

See also Agresti ([2], p. 76). The U model may also be expressed as

$$\theta_{ij} = \theta \quad (i = 1, \dots, r-1; j = 1, \dots, c-1),$$

where

How to cite this paper: Tahata, K., Miyamoto, N. and Tomizawa, S. (2015) Decomposition of Independence Using the Logit Uniform Association Model and Equality of Concordance and Discordance for Two-Way Classifications. *Open Journal of Statistics*, 5, 514-518. <http://dx.doi.org/10.4236/ojs.2015.56054>

$$\theta_{ij} = \frac{p_{ij} p_{i+1, j+1}}{p_{i+1, j} p_{i, j+1}}.$$

Namely this model indicates the constant of the $(r-1)(c-1)$ local odds ratios θ_{ij} defined for adjacent rows and adjacent columns. A special case of the U model obtained by putting $\theta = 1$ is the independence (I) model.

If the I model holds, the correlation coefficient of X and Y equals zero; but the converse does not hold. We are interested in what structure between X and Y is necessary for obtaining the I model, in addition to the correlation coefficient being to zero.

Tomizawa, Miyamoto and Sakurai [3] give the theorem that the I model holds if and only if the Pearson's correlation coefficient ρ for X and Y equals zero and the U model holds.

Tomizawa *et al.* [3] also give the theorem that the I model holds if and only if the Kendall's τ_b equals zero and the U model holds. For τ_b , see Kendall [4] and Agresti ([2], p. 161).

Tahata, Miyamoto and Tomizawa [5] give the theorem that the I model holds if and only if the Spearman's ρ_s equals zero and the U model holds. For ρ_s , see Stuart [6], Kendall and Gibbons ([7], p. 8), and Agresti ([2], p. 164). Also, Tahata and Tomizawa [8] review topics related to the quasi-uniform association model (Goodman [1]), and the decomposition of symmetry into some models for the analysis of *square* contingency tables.

Suppose that the column variable Y is a response variable. Let $L_{j(i)}$ denote the j th cumulative logit within row i ; *i.e.*,

$$L_{j(i)} = \log \left(\frac{G_{i, j+1}^U}{G_{ij}^L} \right),$$

where

$$G_{ij}^L = p_{i1} + \dots + p_{ij},$$

$$G_{i, j+1}^U = p_{i, j+1} + \dots + p_{ic}.$$

The logit uniform association (logit U) model (Agresti [2], p. 122) is defined by

$$L_{j(i+1)} - L_{j(i)} = \beta \quad (i = 1, \dots, r-1; j = 1, \dots, c-1);$$

namely

$$\Theta_{ij} = \beta^* \quad (i = 1, \dots, r-1; j = 1, \dots, c-1),$$

where

$$\Theta_{ij} = \frac{G_{ij}^L G_{i+1, j+1}^U}{G_{i+1, j}^L G_{i, j+1}^U}.$$

Thus the logit U model indicates the constant of the odds ratios for the $(r-1)(c-1) \ 2 \times 2$ tables obtained by taking all pairs of adjacent rows and all dichotomous collapsing of the response (Agresti [2], p. 122). A special case of the logit U model obtained by putting $\beta^* = 1$ (*i.e.*, $\beta = 0$) is the I model. We are now interested in what structure of probabilities $\{p_{ij}\}$ is necessary for obtaining the I model, in addition to the logit U model (instead of the U model).

The purpose of the present paper is to give the decomposition of the I model by using the logit U model (in Section 2).

2. Decomposition of Independence

Let

$$C^* = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} G_{ij}^L G_{i+1, j+1}^U,$$

and

$$D^* = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} G_{i+1,j}^L G_{i,j+1}^U.$$

For a randomly selected pair of observations, 1) $G_{i+1,j+1}^L G_{i,j+1}^U$ is the probability of concordance such that the member that ranks in row $i+1$ rather than in row i also ranks in column $j+1$ or above rather than in column j or below, and 2) $G_{i+1,j}^L G_{i,j+1}^U$ is the probability of discordance such that the member that ranks in row $i+1$ rather than in row i ranks in column j or below rather than in column $j+1$ or above. Therefore C^* and D^* indicate the sum of probabilities of such concordance and those of such discordance, respectively.

We shall consider the model of equality of concordance and discordance (say, CDE model) by

$$C^* = D^*.$$

Then we obtain the following theorem.

Theorem 1. *The I model holds if and only if both the CDE model and the logit U model hold.*

Proof. If the I model holds, i.e., $\{p_{ij} = \mu\alpha_i\beta_j\}$, then

$$C^* = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} G_{ij}^L G_{i+1,j+1}^U = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} \mu^2 \alpha_i \alpha_{i+1} (\beta_1 + \dots + \beta_j) (\beta_{j+1} + \dots + \beta_c),$$

and

$$D^* = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} G_{i+1,j}^L G_{i,j+1}^U = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} \mu^2 \alpha_i \alpha_{i+1} (\beta_1 + \dots + \beta_j) (\beta_{j+1} + \dots + \beta_c).$$

Thus, the CDE model holds. Also, if the I model holds, then the logit U model (with $\beta^* = 1$) holds.

Assuming that both the CDE model and the logit U model hold, then we shall show that the I model holds. Since the logit U model holds, we see

$$G_{ij}^L G_{i+1,j+1}^U = \beta^* G_{i+1,j}^L G_{i,j+1}^U.$$

Thus

$$C^* = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} G_{ij}^L G_{i+1,j+1}^U = \beta^* \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} G_{i+1,j}^L G_{i,j+1}^U = \beta^* D^*.$$

Since the CDE model holds, we obtain $\beta^* = 1$. The proof is completed. \square

Let n_{ij} denote the observed frequency in the (i, j) cell ($i = 1, \dots, r; j = 1, \dots, c$). Assume that a multinomial distribution applies to the $r \times c$ table. Let $G^2(M)$ denote the likelihood ratio chi-squared statistic for testing goodness-of-fit of model M defined by

$$G^2(M) = 2 \sum_{i=1}^r \sum_{j=1}^c n_{ij} \log \left(\frac{n_{ij}}{\hat{m}_{ij}} \right),$$

where \hat{m}_{ij} is the maximum likelihood estimate of expected frequency m_{ij} under the model M. The numbers of degrees of freedom (df) for testing the I, logit U, and CDE models are $(r-1)(c-1)$, $rc-r-c$, and 1, respectively.

3. An Example

The data in **Table 1** are taken directly from Agresti ([2], p. 12), which originally was presented by Grizzle, Starmer and Koch [9]. Four different operations for treating duodenal ulcer patients correspond to removal of various amounts of the stomach. Operation A is drainage and vagotomy, B is 25% resection (antrectomy) and vagotomy, C is 50% resection (hemigastrectomy) and vagotomy, and D is 75% resection. The categories of operation variable have a natural ordering. The dumping severity variable describes the extent of an undesirable potential consequence of the operation. The categories of this variable are also ordered. For these data, the I model fits well with $G^2 = 10.88$ based on $df = 6$. The logit U model also fits these data well with $G^2 = 4.27$

Table 1. Cross-classification of duodenal ulcer patients according to operation and dumping severity.

Operation	Dumping Severity			Total
	None	Slight	Moderate	
A	61	28	7	96
B	68	23	13	104
C	58	40	12	110
D	53	38	16	107
Total	240	129	48	417

Source: Grizzle et al. [9].

based on $df = 5$ (see Agresti ([2], p.123) and Tomizawa [10]). Note that the U model also fits well with $G^2 = 4.59$ based on $df = 5$ (see Agresti ([2], p.81) and Tomizawa [10]).

For testing the hypothesis that the I model holds assuming that the logit U model holds, the difference between the G^2 values for the I model and the logit U model is 6.61 based on $df = 6 - 5 = 1$. Therefore this hypothesis is rejected at the 0.05 level. Hence the logit U model is preferable to the I model for these data.

Also the CDE model fits these data poorly with $G^2 = 5.42$ based on $df = 1$. We see that the rejection of the hypothesis that the I model holds assuming that the logit U model holds is caused by the influence of the lack of structure of the CDE model (*i.e.*, the lack of equality of the sum of probabilities of concordance and those of discordance), because the hypothesis that the I model holds assuming that the logit U model holds is equivalent to the CDE model from **Theorem 1**.

4. Concluding Remarks

When the I model fits the data poorly, **Theorem 1** may be useful for seeing the reason for the poor fit; namely, which of the lack of structure of the CDE model and that of the logit U model influences stronger.

From **Theorem 1** we point out that the hypothesis that the I model holds under the assumption that the logit U model holds is equivalent to the hypothesis that the CDE model holds.

The U model indicates the constant of the $(r-1)(c-1)$ local odds ratios defined for adjacent rows and adjacent columns. On the other hand, the logit U model indicates the constant of the odds ratios for the $(r-1)(c-1) 2 \times 2$ tables obtained by taking all pairs of adjacent rows and all dichotomous collapsing of the response. Thus, when the I model fits the data poorly, if the user wants to see the structure of cumulative probabilities (*i.e.*, the structures of $(r-1)(c-1)$ collapsed 2×2 tables), then **Theorem 1** may be preferable to preceding studies which are described in Section 1.

Acknowledgements

We thank the referee for comments and suggestions.

References

- [1] Goodman, L.A. (1979) Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories. *Journal of the American Statistical Association*, **74**, 537-552.
<http://dx.doi.org/10.1080/01621459.1979.10481650>
- [2] Agresti, A. (1984) *Analysis of Ordinal Categorical Data*. Wiley, New York.
- [3] Tomizawa, S., Miyamoto, N. and Sakurai, M. (2008) Decomposition of Independence Model and Separability of Its Test Statistic for Two-Way Contingency Tables with Ordered Categories. *Advances and Applications in Statistics*, **8**, 209-218.
- [4] Kendall, M.G. (1945) The Treatment of Ties in Ranking Problems. *Biometrika*, **33**, 239-251.
<http://dx.doi.org/10.1093/biomet/33.3.239>
- [5] Tahata, K., Miyamoto, N. and Tomizawa, S. (2008) Decomposition of Independence Using Pearson, Kendall and Spearman's Correlations and Association Model for Two-Way Classifications. *Far East Journal of Theoretical Statistics*, **25**, 273-283.

- [6] Stuart, A. (1963) Calculation of Spearman's Rho for Ordered Two-Way Classifications. *The American Statistician*, **17**, 23-24.
- [7] Kendall, M. and Gibbons, J.D. (1990) Rank Correlation Methods. 5th Edition, Edward Arnold, London.
- [8] Tahata, K. and Tomizawa, S. (2014) Symmetry and Asymmetry Models and Decompositions of Models for Contingency Tables. *SUT Journal of Mathematics*, **50**, 131-165.
- [9] Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969) Analysis of Categorical Data by Linear Models. *Biometrics*, **25**, 489-504. <http://dx.doi.org/10.2307/2528901>
- [10] Tomizawa, S. (1992) More Parsimonious Linear-by-Linear Association Model in the Analysis of Cross-Classifications Having Ordered Categories. *Biometrical Journal*, **34**, 129-140. <http://dx.doi.org/10.1002/bimj.4710340202>