International Journal of Statistics and Probability; Vol. 7, No. 4; July 2018
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Extended Marginal Homogeneity Model Based on Complementary Log-Log Transform for Square Tables

Yusuke Saigusa[1], Tomohisa Maruyama[2], Kouji Tahata[2] & Sadao Tomizawa[2]

[1] Department of Biostatistics, Yokohama City University School of Medicine, Japan

[2] Department of Information Sciences, Faculty of Science and Technology, Tokyo University of Science, Japan

Correspondence: Yusuke Saigusa, Department of Biostatistics, Yokohama City University School of Medicine, Yokohama, Kanagawa 236–0004, Japan.

Received: April 2, 2018 Accepted: April 17, 2018 Online Published: May 28, 2018

doi:10.5539/ijsp.v7n4p27 URL: https://doi.org/10.5539/ijsp.v7n4p27

Abstract

For square contingency tables with the same ordinal row and column classifications, McCullagh (1977) gave the marginal cumulative logistic model, which is an extension of the marginal homogeneity (MH) model using the logit transform. The present paper proposes a different extension of the MH model using the complementary log-log transform. In addition, the present paper gives the theorem that the MH model is equivalent to the proposed model and the equality of row and column marginal means holding simultaneously. In data analysis, if the MH model fits the data poorly, the theorem may be useful for seeing the reason for the poor fit. As example, the occupational status data for British father-son pairs are analyzed.

Keywords: decomposition, mean equality, logit transform

1. Introduction

Consider a square contingency table with the same ordinal row and column classifications. In the data in Table 1 taken from Bishop, Fienberg & Holland (1975, p.100), each observation is a pair of father's occupational status with his son's occupational status. For such data, statistical independence between the row and column classification generally does not hold due to concentration of observations on main diagonal cells. Instead of independence, we are interested in whether there is a structure of symmetry in the table. For example, Stuart (1955) gave the marginal homogeneity (MH) model which states the row marginal distribution is identical to the column marginal distribution. It is known that the MH model is expressed as the equality of marginal cumulative probabilities of row and column. For the data in Table 1, the MH model indicates the probability that a father's status is i equals the probability that his son's status is also i for any category i.

In data analysis, when the MH model fits the data poorly, many statisticians may be interested in a comparison of the two marginal distributions of row and column variables, say X and Y. One of such analyses is inferring whether X tends to be stochastically less than Y or vice versa. We are especially interested in applying the extension of the MH model, for example, the marginal cumulative logistic (ML) model (McCullagh, 1977; Agresti, 1984, p.205) based on the logit transform. The ML model states that one marginal distribution is a location shift of the other marginal distribution on a logistic scale. If the ML model fits the data poorly, we are then interested in other extension of the MH model based on the complementary log-log transform rather than logit transform.

Miyamoto, Niibe & Tomizawa (2005) gave the theorem that the MH model holds if and only if the ML model and the equality of row and column marginal means hold simultaneously. We refer to such relation as a decomposition of model (i.e., the MH model is decomposed into the ML model and the equality of row and column marginal means). Also, see Tahata & Tomizawa (2008) and Kurakami, Tahata & Tomizawa (2013) for the decompositions of the MH model. We are interested in whether the decomposition with the ML model replaced by the proposed model holds or not. When the MH model fits the data poorly, it may be useful for seeing the reason for the poor fit of it.

In this paper, Section 2 proposes a new model which is an extension of the MH model based on the complementary log-log transform. Section 3 gives the decomposition of the MH model using the proposed model. Section 4 refers to the goodness-of-fit test. Section 5 analyzes the father's and his son's occupational mobility data in Britain. We show that the new model and decomposition are useful for inferring relationships between marginal distributions with the example.

2. Models

For an $r \times r$ square contingency table with ordered categories, let p_{ij} denote the probability that an observation will fall in the ith row and jth column of the table for $i = 1, \ldots, r$; $j = 1, \ldots, r$. The MH model is defined by

$$p_{i\cdot} = p_{\cdot i} \quad (i = 1, \ldots, r),$$

where $p_{i\cdot} = \sum_{t=1}^{r} p_{it}$ and $p_{\cdot i} = \sum_{s=1}^{r} p_{si}$ (Stuart, 1955; Tahata & Tomizawa, 2014). This model indicates the structure that satisfies the identity of marginal distributions of row and column. Let F_i^X and F_i^Y denote the marginal cumulative probability of X and Y, respectively; namely $F_i^X = \sum_{s=1}^{i} p_{s\cdot}$ and $F_i^Y = \sum_{t=1}^{i} p_{\cdot t}$ for $i = 1, \ldots, r-1$. The MH model may also be expressed as

$$F_i^X = F_i^Y \quad (i = 1, \ldots, r-1).$$

Let L_i^X and L_i^Y denote the marginal cumulative logit transforms of X and Y, respectively; namely

$$L_i^X = \log\left(\frac{F_i^X}{1 - F_i^X}\right), \quad L_i^Y = \log\left(\frac{F_i^Y}{1 - F_i^Y}\right) \quad (i = 1, \ldots, r-1).$$

The MH model may further be expressed as

$$L_i^X = L_i^Y \quad (i = 1, \ldots, r-1).$$

The ML model (McCullagh, 1977) is defined by

$$L_i^X = L_i^Y + \delta \quad (i = 1, \ldots, r-1),$$

where the parameter δ is unspecified. The ML model is one of the extensions of the MH model. This model indicates that the odds that X is i or below instead of $i+1$ or above, is $\exp(\delta)$ times higher than the odds that Y is i or below instead of $i+1$ or above, for $i = 1, \ldots, r-1$. Therefore this model states one marginal distribution is a location shift of the other marginal distribution on a logistic scale.

Let C_i^X and C_i^Y denote the marginal cumulative complementary log-log transforms of X and Y, respectively; namely

$$C_i^X = \log\left(-\log\left(1 - F_i^X\right)\right), \quad C_i^Y = \log\left(-\log\left(1 - F_i^Y\right)\right) \quad (i = 1, \ldots, r-1).$$

The MH model may be expressed as

$$C_i^X = C_i^Y \quad (i = 1, \ldots, r-1).$$

We shall consider now the marginal cumulative complementary log-log (MCL) model which is defined by

$$C_i^X = C_i^Y + \log\Delta \quad (i = 1, \ldots, r-1),$$

where Δ is unspecified. This model indicates that the probability that X is $i+1$ or above, is equal to the probability that Y is $i+1$ or above to the power of Δ, for $i = 1, \ldots, r-1$. Thus this model states one marginal distribution is a location shift of the other marginal distribution on a complementary log-log scale. Note that if $\Delta = 1$, then we have the MH model. We see, under the MCL model, $\Delta > 1$ is equivalent to $F_i^X > F_i^Y$ and $\Delta < 1$ is equivalent to $F_i^X < F_i^Y$. Therefore the parameter Δ in the MCL model reflects the degree of inhomogeneity between $\{F_i^X\}$ and $\{F_i^Y\}$.

3. Decomposition

Consider the specified scores $\{u_k\}$ may be assigned to both rows and columns satisfying $u_1 \leq u_2 \leq \cdots \leq u_r$ or $u_1 \geq u_2 \geq \cdots \geq u_r$, where at least one strict inequality holds. Using the function $g(k)$ which is $g(k) = u_k$ for $k = 1, \ldots, r$, consider the marginal mean equality (ME) model defined by

$$E(g(X)) = E(g(Y)),$$

where $E(g(X)) = \sum_{i=1}^{r} g(i)p_{i\cdot}$ and $E(g(Y)) = \sum_{i=1}^{r} g(i)p_{\cdot i}$.

We now obtain the following theorem.

Theorem 1. *The MH model holds if and only if both the MCL and ME models hold.*

proof. If the MH model holds, then the MCL and ME models hold. We assume that both the MCL and ME models hold, and then we show that the MH model holds. For $u_1 \leq u_2 \leq \cdots \leq u_r$ (or $u_1 \geq u_2 \geq \cdots \geq u_r$), we have

$$E(g(X)) = \sum_{i=1}^{r} g(i)p_{i\cdot} = g(1) + \sum_{k=1}^{r-1} d_k \left(1 - F_k^X\right),$$

where

$$d_k = g(k+1) - g(k).$$

Similarly, we have

$$E(g(Y)) = g(1) + \sum_{k=1}^{r-1} d_k \left(1 - F_k^Y\right).$$

Since the ME and MCL models hold, we have

$$\sum_{k=1}^{r-1} d_k \left(1 - F_k^X\right) = \sum_{k=1}^{r-1} d_k \left(1 - F_k^Y\right), \tag{1}$$

and

$$\sum_{k=1}^{r-1} d_k \left(1 - F_k^X\right) = \sum_{k=1}^{r-1} d_k \left(1 - F_k^Y\right)^{\Delta}. \tag{2}$$

Equations (1) and (2) lead to

$$\sum_{k=1}^{r-1} d_k \left(1 - F_k^Y\right) = \sum_{k=1}^{r-1} d_k \left(1 - F_k^Y\right)^{\Delta}.$$

Thus we obtain $\Delta = 1$, i.e., the MH model holds because $d_k \geq 0$ (or $d_k \leq 0$) for all $k = 1, \ldots, r-1$, with at least one of the $\{d_k\}$ being not equal to zero. The proof is completed.

4. Goodness-of-fit Test

Let n_{ij} denote the observed frequency in the ith row and jth column of the $r \times r$ table with $n = \sum\sum n_{ij}$, and let m_{ij} denote the corresponding expected frequency for $i = 1, \ldots, r;\ j = 1, \ldots, r$. We assume that a multinomial distribution applies to the table. The maximum likelihood estimates (MLEs) of expected frequencies under each model can be obtained using the Newton-Raphson method in the log-likelihood equation (see Appendix for the log-likelihood equation). The likelihood ratio chi-squared statistic for testing the goodness-of-fit of model M is given by

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

where $\hat{m}_{ij}$ is the MLE of m_{ij} under the model. The numbers of degrees of freedom (df) of statistics for testing the goodness-of-fit of the MH, ML, MCL, and ME models are $r-1$, $r-2$, $r-2$, and 1, respectively. Consider two nested models, say M_1 and M_2, such that if model M_1 holds, then model M_2 holds. For testing the goodness-of-fit of model M_1 assuming that model M_2 holds, the conditional likelihood ratio statistic is given by $G^2(\mathrm{M}_1|\mathrm{M}_2) = G^2(\mathrm{M}_1) - G^2(\mathrm{M}_2)$. The number of df for the conditional test is the difference between the numbers of df for the models M_1 and M_2.

5. Example

Consider the data in Table 1, relating the father's and his son's occupational status categories for a British sample again. The smaller category number means the higher status. We analyze the data using the new model and decomposition of the MH model.

Table 2 gives the values of likelihood ratio statistic G^2 for testing the goodness-of-fit of models. We set $u_k = k$ for $k = 1, \ldots, 5$. The MH, ML and ME models fit the data poorly (G^2(MH) = 32.80 with 4 df; G^2(ML) = 9.75 with 3 df; G^2(ME) = 20.28 with 1 df). The MCL model fits the data well (G^2(MCL) = 4.26 with 3 df). Using Theorem 1 which is the decomposition of the MH model into the MCL and ME models, we shall consider the reason why the MH model fits the data poorly. According to Theorem 1 and Table 2, the poor fit of the MH model is caused by the influence of the lack of structure of the ME model rather than the MCL model. Note that, using the decomposition of the MH model into the ML and ME models, it is difficult to consider the reason for the poor fit of the MH model because both the ML and ME models fit the data poorly.

Since the MCL model which is implied by the MH model fits well, we can test the goodness-of-fit of the MH model under the assumption that the MCL model holds, i.e., the hypothesis that $\Delta = 1$ under the assumption. The difference between the G^2 values for the MH and MCL models is G^2(MH|MCL) = G^2(MH) − G^2(MCL) = 28.54 with $4 - 3 = 1$ df, and thus the hypothesis that $\Delta = 1$ is rejected at the 0.05 significance level. It shows strong evidence of $\Delta \neq 1$ in the MCL model. Therefore the MCL model is preferable to the MH model for the data. Under the MCL model, the MLE of Δ is $\hat{\Delta} = 1.13$.

Namely, under the MCL model, the probability that the status category for father in a pair is $i+1$ or above, is estimated to be equal to the probability that the status category for son in the pair is $i+1$ or above to the power of 1.13, for $i = 1, \dots, 4$. Since $\hat{\Delta} > 1$, under the MCL model, $\hat{F}_i^X > \hat{F}_i^Y$, where $\hat{F}_i^X$ and $\hat{F}_i^Y$ are MLEs of the marginal cumulative probabilities of X and Y for $i = 1, \dots, 4$. Therefore the distribution of the status category for the son tends to be stochastically higher than that for his father.

Acknowledgements

The authors would like to thank two referees and the editor for their helpful comments.

References

Agresti, A. (1984). *Analysis of ordinal categorical data*. Wiley, New York. https://doi.org/10.1002/bimj.4710290113

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. MIT Press, Cambridge. https://doi.org/10.1007/978-0-387-72806-3

Kurakami, H., Tahata, K., & Tomizawa, S. (2013). Generalized marginal cumulative logistic model for multi-way contingency tables. *SUT Journal of Mathematics, 49*, 19–32. https://doi.org/10.20604/00000925

McCullagh, P. (1977). A logistic model for paired comparisons with ordered categorical data. *Biometrika, 64*, 449–453. https://doi.org/10.1093/biomet/64.3.449

Miyamoto, N., Niibe, K., & Tomizawa, S. (2005). Decompositions of marginal homogeneity model using cumulative logistic models for square contingency tables with ordered categories. *Austrian Journal of Statistics, 34*, 361–373. https://doi.org/10.17713/ajs.v34i4.424

Stuart, A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika, 42*, 412–416. https://doi.org/10.1093/biomet/42.3-4.412

Tahata, K., & Tomizawa, S. (2008). Generalized marginal homogeneity model and its relation to marginal equimoments for square contingency tables with ordered categories. *Advances in Data Analysis and Classification, 2*, 295–311. https://doi.org/10.1007/s11634-008-0028-1

Tahata, K., & Tomizawa, S. (2014). Symmetry and asymmetry models and decompositions of models for contingency tables. *SUT Journal of Mathematics, 50*, 131–165. https://doi.org/10.20604/00000822

Table 1. Occupational status for British father-son pairs (Bishop *et al.*, 1975, p.100)

	Son's status					
Father's status	1	2	3	4	5	Total
1	50	45	8	18	8	129
	(50.25)	(40.88)	(7.83)	(17.62)	(7.65)	(124.24)
2	28	174	84	154	55	495
	(31.07)	(172.82)	(90.59)	(166.05)	(57.80)	(518.33)
3	11	78	110	223	96	518
	(11.30)	(72.29)	(110.07)	(223.10)	(93.79)	(510.56)
4	14	150	185	714	447	1510
	(14.39)	(139.05)	(185.16)	(714.46)	(436.78)	(1489.84)
5	3	42	72	320	411	848
	(3.16)	(39.86)	(73.91)	(328.43)	(411.67)	(857.03)
Total	106	489	459	1429	1017	3500
	(110.17)	(464.90)	(467.57)	(1449.66)	(1007.70)	(3500.00)

Note: The parenthesized values are the MLEs of expected frequencies under the MCL model.

Table 2. Likelihood ratio chi-square values G^2 for models applied to the data in Table 1

Models	df	G^2
MH	4	32.80^*
ML	3	9.75^*
MCL	3	4.26
ME	1	20.28^*

Note: u_k for the ME model is integer score. * means significant at the 0.05 level.

Appendix

We consider the MLEs of the expected frequencies $\{m_{ij}\}$ under the MCL model. Those under the MH, ML and ME models can be obtained in the similar manner, although those are omitted here.

To obtain MLEs under the MCL model, we must maximize the Lagrangian

$$L = \sum_{i=1}^{r}\sum_{j=1}^{r} n_{ij}\log p_{ij} - \lambda\left(\sum_{i=1}^{r}\sum_{j=1}^{r} p_{ij} - 1\right) - \sum_{i=1}^{r-1}\mu_i\left(\log\left(1 - F_i^X\right) - \Delta\log\left(1 - F_i^Y\right)\right)$$

with respect to $\{p_{ij}\}$, λ, $\{\mu_i\}$, and Δ. Setting the partial derivatives of L equal to zero, we obtain the equations

$$p_{ij} = n_{ij}\left\{n + \sum_{k=1}^{r-1}\mu_k\left(\frac{F_k^X - I(i \le k)}{1 - F_k^X} - \frac{\Delta\left(F_k^Y - I(j \le k)\right)}{1 - F_k^Y}\right)\right\}^{-1} \quad (i = 1, \dots, r;\ j = 1, \dots, r),$$

as well as

$$1 - F_i^X = \left(1 - F_i^Y\right)^{\Delta} \quad (i = 1, \dots, r - 1),$$

and

$$\sum_{i=1}^{r-1}\mu_i\log\left(1 - F_i^Y\right) = 0,$$

where $I(\cdot)$ is the indicator function. Using the Newton-Raphson method, we can solve the equations with respect to $\{p_{ij}\}$, $\{\mu_i\}$ and Δ. Then we can obtain the MLEs of $\{m_{ij}\}$ and Δ under the MCL model.

Copyrights