Hacettepe Journal of Mathematics and Statistics
Volume 39 (1) (2010), 75 – 80

ESTIMATION OF THE CORRELATION COEFFICIENT FOR TRIANGULAR CONTINGENCY TABLES UNDER THE BIVARIATE NORMAL DISTRIBUTION

Serpil Aktaş*

Received 18 : 09 : 2007 : Accepted 18 : 11 : 2009

Abstract

Triangular contingency tables are a special class of incomplete contingency tables. Association and independence models are used to analyze such tables. Association models can be described in terms of the association parameters for the analysis of triangular contingency tables having ordered categories. The aim of this study is to show the relation between the association parameters of the uniform association model and the sample correlation coefficient under the structural zeros. For this purpose, a simulation study based on random contingency tables containing structural zeros is performed. Association parameters are estimated under the uniform association models. The sample correlation coefficients are computed using these parameter estimates and compared with the population correlation coefficients. It is shown that by using the association parameter estimates under the uniform association model, better estimates can be achieved for the population correlation coefficient in the case of structural zeros.

Keywords: Triangular contingency tables, Uniform association model, Bivariate normal distribution.

2000 AMS Classification: 62 H 17.

*Department of Statistics, Hacettepe University, 06800 Beytepe, Ankara, Turkey.
E-mail: spxl@hacettepe.edu.tr

1. Introduction

Statistical methods for qualitative data have received considerable attention in recent years. The variable of interest is often a categorical variable in the same way as in social sciences, biomedical studies and behavioral sciences. There are many statistical procedures which can be used for the analysis of categorical data. A special class of incomplete contingency tables is the triangular contingency tables, which contain structural zeros in one or more cells above or below their main diagonals.

Triangular contingency tables were first analyzed in Goodman [4] by partitioning the table into a set of rectangular subtables, each of which can be analyzed in an elementary way. Bishop and Fienberg [2] discussed these kinds of table with the classical example of disability of stroke patients. Altham [1], Mantel [9] and Bishop *et al.* [3] also discussed the quasi-independence model. Goodman [4] introduced various tests of the quasi-independence (QI) model against alternative hypothesis of positive or negative quasi-dependence. The QI model omits the ordinal nature of the row and column variables. Therefore, association models have been suggested for analyzing ordinal contingency tables.

In this paper, ordinal triangular tables are analyzed through the association model, rather than the QI model. Goodman[5] defined the sample correlation coefficient in terms of the association parameters which is estimated under the uniform association model. In this paper, we investigate whether this relation is still valid when the data consists of structural zeros. But only triangular contingency table forms are considered, as a special class of incomplete contingency tables.

2. Triangular tables

A contingency table is a tabular representation of categorical data. A contingency table usually shows the frequencies for particular combinations of the values of two discrete random variables X and Y. Each cell in the table represents a mutually exclusive combination of X-Y values. We consider $R \times R$ square tables, where the row and the column categories are ordinal, numbered from 1 to R, and π_{ij} will denote the probability that an observation falls in the ith row and jth column of the table.

Sarkar [10] defined four types of triangular contingency table under the following conditions: An *upper-right* (*left*) *triangular table* (URT) ((ULT)) is described by the condition that $\pi_{ij} = 0$ for $i > j$ $(i + j > R + 1)$, and a *lower-left* (*right*) *triangular table* (LLT) ((LRT)) for $i < j$, $(i + j < R + 1)$. Any of these tables can be transformed to the other by interchanging the row and column variables and/or reversing the category ordering.

For the URT tables there will not be any observations for $i < j$. An empty cell in which the observations are impossible is called a *structural zero*. A structural zero is not an observation, and is not part of the data. Contingency tables with structural zeros are called incomplete tables. For contingency tables with the triangular structure, it is obvious that the null hypothesis of independence cannot be expected to fit. Due to the incompleteness of the table, the independence between the row and the column variables are tested by the quasi-independence model first proposed by Goodman [5]. The QI model in a URT table is given by

$$\pi_{ij} = \begin{cases} \alpha_i \beta_j & i \leqslant j, \\ 0 & i > j, \end{cases} \tag{2.1}$$

where $\alpha_i > 0$ and $\beta_j > 0$ are positive constants for $i = 1, \ldots, R$, and $j = 1, \ldots, R$. For testing the QI model, Goodman [6], Bishop and Fienberg [2] used usual chi-squared tests

based on the Pearson and the log-likelihood ratio statistics. For the analysis of two-way cross classification table having ordered categories when the row and the column variables are ordinal, the quasi-independence model does not take into account the ordinal nature of the tables. Association models usually give better results.

Goodman [7] considered various kinds of association model. Suppose that both the row and column variables of a two-dimensional table are ordinal, with the row variable denoted by X and the column variable by Y. We assume that the scores $\{u_i\}$ and $\{v_j\}$ are assigned to the rows and columns, respectively, where $u_1 < u_2 < \cdots < u_R$, $v_1 < v_2 < \cdots < v_R$. A simple log-linear model that uses the ordinal information, but that has only one more parameter than the usual independence model is given by

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \beta^{XY}(u_i - \bar{u})(v - \bar{v}), \tag{2.2}$$

where $\sum_{i=1}^R \lambda_i^X = \sum_{i=1}^R \lambda_j^Y$ and $\bar{u}$ and $\bar{v}$ are the arithmetic means, respectively.

This model is referred to as the *linear-by-linear association model*, and requires the assignment of the scores. The parameter β in Model 2.2, that describes the association between X and Y, can be interpreted as the common value of the local log-odds ratio. The model effectively means that all adjacent odds ratios have the same value. Its sign is related to the direction of the association; for example, if higher values of X are associated with higher values of Y, the sign of β will be positive. When $\beta = 0$, it indicates independence. The independence model is the special case $\beta^{XY} = 0$.

Goodman [7] suggested a model for the special case $\{u_i = i\}$, $\{v_i = j\}$, in which the local odds ratio θ_{ij} is uniformly $\exp(\beta)$ for adjacent rows i and $i+1$, and adjacent columns j and $j+1$. Goodman [7] referred to this special case as the uniform association (UA) model.

3. Bivariate normal distribution

Let X and Y be random variables that have the joint probability density function

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\Big\{ -\frac{1}{2(1-\rho^2)}\Big[(\frac{x-\mu_x}{\sigma_x})^2 - 2\rho(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y}) + (\frac{y-\mu_y}{\sigma_y})^2\Big]\Big\}$$
$$-\infty < x < +\infty,\ -\infty < y < +\infty,$$

where ρ is the correlation coefficient between X and Y and the correlation is the linear association between the two random variables X and Y. The value of the coefficient ranges from -1 to 1. If ρ is 0, X and Y are said to be uncorrelated, with no linear association between X and Y. In the formula, the standard deviations σ_x and σ_y are positive constants, but the means μ_x and μ_y do not have to be positive constants. The random variables X and Y are said to have the *bivariate normal distribution*.

For normalized variables $z_x = \frac{(x-\mu_x)}{\sigma_x}$ and $z_y = \frac{(y-\mu_y)}{\sigma_y}$ the bivariate normal probability density function becomes,

$$f(z_1,z_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp(-\frac{1}{2(1-\rho^2)})(z_1^2 - 2\rho z_1 z_2 + z_2^2).$$

As mentioned above, the parameter β in the uniform association model gives the association between the row and the column variables, and Goodman [5] defined the association parameter in terms of the correlation as,

$$\widehat{\beta} = \frac{\hat{\rho}}{(1-\hat{\rho}^2)}. \tag{3.1}$$

Hence, the association parameter is an estimator of the correlation coefficient. Association models can be used to fit a cross-classification table, when the row and column classifications arise from underlying continuous random variables having a bivariate normal distribution.

4. Simulation study

A simulation study was performed to generate pseudo $R \times R$ triangular contingency tables having ordered categories. Random samples of size N were generated from a bivariate normal distribution with correlation ρ [8]. Samples drawn from the bivariate normal distribution with a known correlation coefficient were transformed into contingency tables of equal-interval frequency. After the discretization of the data, simulation parameters were set as: sample size $n = 100, 250, 500, 1000$; correlation coefficient $\rho = 0.0, 0.2, 0.4$; and dimension $R = 4, 5, 6, 7$, and 8. In order to generate uncorrelated data, correlation coefficients were taken to be small. After 500 replications in each combination, 30,000 URT tables were generated. Table 1 shows an example of 5×5 URT table.

Table 1. A 5×5 upper right triangular table

n_{ij}	1	2	3	4	5
1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}
2	-	n_{22}	n_{23}	n_{24}	n_{25}
3	-	-	n_{33}	n_{34}	n_{35}
4	-	-	-	n_{44}	n_{45}
5	-	-	-	-	n_{55}

As an illustrative example, we show below the design matrix of a uniform association model for a 5×5 URT table.

$$\log \begin{pmatrix} m_{11} \\ m_{12} \\ m_{13} \\ m_{14} \\ m_{15} \\ m_{21} \\ m_{22} \\ m_{23} \\ m_{24} \\ m_{25} \\ m_{31} \\ m_{32} \\ m_{33} \\ m_{34} \\ m_{35} \\ m_{41} \\ m_{42} \\ m_{43} \\ m_{44} \\ m_{45} \\ m_{51} \\ m_{52} \\ m_{53} \\ m_{54} \\ m_{55} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 3 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 4 \\ 1 & 1 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 5 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 2 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 4 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 6 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 8 \\ 1 & 0 & 1 & 0 & 0 & -1 & -1 & -1 & -1 & 10 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 3 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 6 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 9 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 12 \\ 1 & 0 & 0 & 1 & 0 & -1 & -1 & -1 & -1 & 15 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 4 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 8 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 12 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 16 \\ 1 & 0 & 0 & 0 & 1 & -1 & -1 & -1 & -1 & 20 \\ 1 & -1 & -1 & -1 & -1 & 1 & 0 & 0 & 0 & 5 \\ 1 & -1 & -1 & -1 & -1 & 0 & 1 & 0 & 0 & 10 \\ 1 & -1 & -1 & -1 & -1 & 0 & 0 & 1 & 0 & 15 \\ 1 & -1 & -1 & -1 & -1 & 0 & 0 & 0 & 1 & 20 \\ 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & 25 \end{pmatrix} \begin{pmatrix} \mu \\ \lambda_1^X \\ \lambda_2^X \\ \lambda_3^X \\ \lambda_4^X \\ \lambda_1^Y \\ \lambda_2^Y \\ \lambda_3^Y \\ \lambda_4^Y \\ \beta \end{pmatrix}$$

A uniform association model was applied to the tables generated by the simulation, and goodness of fit statistics and P-values were computed. Among the generated tables, the ones for which the null hypothesis is not rejected were selected ($P > 0.05$). Parameter values were estimated under the structural zero frequencies only for those triangular contingency tables which fit the UA model. By solving Equation 3.1 using the estimated

β parameters, Pearson correlation coefficients were estimated and are given in Table 2 with $\widehat{\beta}$ values.

Table 2. Estimated correlation coefficients and values of $\widehat{\beta}$ for the generated tables

	n	100		250		500		1000	
	R	$\widehat{\beta}$	$\hat{\rho}$	$\widehat{\beta}$	$\hat{\rho}$	$\widehat{\beta}$	$\hat{\rho}$	$\widehat{\beta}$	$\hat{\rho}$
$\rho = 0.1$	4×4	0.5004	0.4144	0.5184	0.4248	0.5171	0.4241	0.5424	0.4382
	5×5	0.3409	0.3085	0.3738	0.3325	0.4066	0.3553	0.4342	0.3789
	6×6	0.2418	0.2291	0.2768	0.2583	0.3058	0.2816	0.3378	0.3061
	7×7	0.1810	0.1754	0.1997	0.1923	0.2230	0.2129	0.2557	0.2409
	8×8	0.1268	0.1248	0.1493	0.1461	0.1677	0.1632	0.1967	0.1896
$\rho = 0.2$	4×4	0.5272	0.4298	0.5099	0.4200	0.5218	0.4268	0.5529	0.4439
	5×5	0.3213	0.2936	0.3728	0.3318	0.4127	0.3594	0.4368	0.3753
	6×6	0.2456	0.2323	0.2815	0.2622	0.3129	0.2871	0.3337	0.3031
	7×7	0.1809	0.1753	0.1992	0.1919	0.2262	0.2157	0.2417	0.2290
	8×8	0.1286	0.1265	0.1553	0.1517	0.1692	0.1646	0.1927	0.1860
$\rho = 0.4$	4×4	0.4959	0.4188	0.5179	0.4246	0.5404	0.4371	0.5543	0.4447
	5×5	0.3426	0.3097	0.3739	0.3326	0.4291	0.3703	0.4428	0.3791
	6×6	0.2362	0.2243	0.2753	0.2571	0.3034	0.2797	0.3222	0.2943
	7×7	0.1713	0.1665	0.2063	0.1982	0.2254	0.2150	0.2536	0.2391
	8×8	0.1282	0.1262	0.1517	0.1484	0.1705	0.1658	0.1872	0.1811

From Table 2, we can state in general that the results are not affected by the sample size. For $\rho = 0.1$, as the dimension increases, $\hat{\rho}$ approaches to ρ. On the contrary, for $\rho = 0.4$, as the dimension increases, $\hat{\rho}$ differs from ρ. For $\rho = 0.2$, $R = 6$ and $R = 7$ give reasonable estimates. The combination $R = 4$ and $\rho = 0.1$ does not give realistic estimates for any sample size.

5. Conclusions

The QI model disregards the ordinal nature of the row and the column variables. In order to avoid this problem, we applied the UA model to the generated tables, taking integer scores and interpreted the QI model in terms of the ordinal association. Table 2 gives the maximum likelihood estimation of the β parameters for 500 replications. Using Equation (3.1), we have found the correlation coefficients. It can be seen that, in some cases, the estimated correlation coefficients obtained from the association parameters converge to the actual correlations. The simulation results are not affected by the sample size. We note that as the dimension increases, the estimates decrease and tend to ρ. It has been shown that the relation between β and ρ is still valid for some cases when the data contains structural zeros. These results are valid only for triangular contingency tables, not for arbitrary contingency table with structural zeros.

References

[1] Altham, P. M. *Quasi-independent triangular contingency tables*, Biometrics **31**, 233–238, 1975.

[2] Bishop, Y. M. M. and Fienberg. S. E. *Incomplete two-dimensional contingency tables*, Biometrics **28**, 177–202, 1969.

[3] Bishop, Y. M, M., Fienberg, S. E. and Holland. P. W. *Discrete Multivariate Analysis* (Cambridge, MA: MIT Press, 1975).

[4] Goodman, L. A. *On quasi-independence in triangular contingency tables*, Biometrics **35**, 651–655, 1979.

[5] Goodman, L. A. *The analysis of cross-classified data: Independence, quasi-independence and interactions in contingency tables with or without missing entries*, J. A. S. A. **63**, 1091–1131, 1968.

[6] Goodman, L. A. *Association models and the bivariate normal distribution in the analysis of cross-classifications having ordered categories*, Biometrika **68**, 347–355, 1981.

[7] Goodman, L. A. *The Analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models,and asymmetry models for contingency tables with or without missing entries*, The Annals of Statistics **13** (1), 10–69, 1985.

[8] Goodman, L. A. *On quasi-independence and quasi-dependence in contingency tables, with special reference to ordinal triangular contingency tables*, J. A. S. A. **89**, 1059–1063, 1994.

[9] Mantel, N. *Incomplete contingency tables*, Biometrics **26**, 291–304, 1970.

[10] Sarkar, S. K. *Quasi-independence in ordinal triangular contingency tables*, J. A. S. A. **84**, 592–597, 1989.