The Korean Communications
in Statistics Vol. 12 No. 2, 2005
pp. 521-529

Bootstrap Method for Row and Column Effects Model

Hyeong Chul Jeong[1]

Abstract

In this paper, we consider a bootstrap method to the "row and column effects model" (RC model) to analyze a contingency table with ordered variables. We propose a bootstrap procedure for testing of independence, equality of intervals, and goodness of fit in the RC model. A real data example is included.

Keywords : Bootstrap, RC model, Independence test, Equality of interval test, Goodness of fit test

1. Introduction

In recent years, the analysis of association among ordinal categorical data has received considerable attention. There are several advantages to be gained from using ordinal models which use information on ordering instead of the standard procedures appropriate for nominal categorical data. For ordinal categorical data, a greater variety of models exist which are more parsimonious and have simpler interpretations than the nominal methods (Davis, 1988). A list of pertinent references includes Agresti (1984), Goodman (1979, 1986), Gilula and Ritov (1990) and Jeong, Jhun and Kim (2005).

Consider an $r \times c$ contingency table having ordinal variables. Let X be the row category variable and let Y be the column category variable. The log-linear models treat all classifications as nominal, in the sense that parameter estimates are invariant to orderings of categories. These models ignore important information that reflects the orderings, such as "the X variable increases as the Y variable increases or decreases". Goodman (1979, 1981a, 1981b) provided a "row and column effects model" (RC model) to analyze a contingency table with ordered variables. The RC model treats the row and column scores as parameters to be estimated from the data. Agresti (1984, 1990) calls it the log-multiplicative or log-bilinear model because the log expected frequency is a multiplicative (rather than linear) function of the model parameters. The RC model has a nonlinear property for parameters because the log

[1] Assistant Professor, Department of Applied Statistics, University of Suwon, Kyunggi, 445-743, Korea. E-mail : jhc@suwon.ac.kr

expected frequency is a multiplicative function of the parameters. Therefore, there are many difficulties in using the conventional statistical inference in the RC model. And it is very difficult to identify the distribution of the row and column scores with the RC model. Therefore, a bootstrap method for the RC model is considered to overcome the above problems. Jeong *et al.* (2005) showed that the bootstrap method for the two-way ordinal contingency tables outperforms the other conventional methods for the test of independence, at least when the sample size is small and the dimension of the contingency table is large. In this paper, we consider the various statistical inferences related to the RC model using the bootstrap method. The methods discussed in this paper are also applied to a real data set.

The remainder of this paper is organized as follows. Section 2 presents some preliminaries about the RC model. In section 3, we propose a bootstrap procedure for testing of independence, testing of equality of intervals, and testing of goodness of fit in the RC model. A real data example is presented in section 4. Finally, we give concluding remarks in section 5.

2. Row and Column Effects Model

For an $r \times c$ contingency table, a general form for the association model is

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + u_i v_j \tag{2.1}$$

where $\sum_i \lambda_i^X = \sum_j \lambda_j^Y = 0$, u_i represents row scores of X variables, and v_j denotes column scores of Y variables. In this association model, the log-odds ratio is

$$\log \theta_{ij} = \log \frac{m_{ij} m_{i+1,j+1}}{m_{i,j+1} m_{i+1,j}} = (u_{i+1} - u_i)(v_{j+1} - v_j).$$

Model (2.1) is the "linear-by-linear association model" when $u_i v_j = \beta u_i^* v_j^*$ with u_i and v_j being fixed, strictly monotone scores (Agresti, 1990).

The RC model is written as:

$$\log m_{ij} = \mu + \lambda_i^X + \lambda_j^Y + \beta u_i v_j \tag{2.2}$$

Here the constraints are $\sum_i u_i p_{i+} = \sum_j v_j p_{+j} = 0$ and $\sum_i u_i^2 p_{i+} = \sum_j v_j^2 p_{+j} = 1$, where p_{i+} and p_{+j} denote the row and column sample marginal distributions, respectively. Under the constraints, the estimated association parameter $\hat{\beta}$ can be interpreted as a correlation coefficient. Goodman (1981b) discussed the fact that the RC model for discrete variables has a form which is similar to the bivariate normal density for continuous variables. If (X, Y) have a standardized bivariate normal distribution, then

$$\begin{aligned} f(x,y) &= (2\pi\sqrt{1-\rho^2})^{-1} \exp\left[\frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right] \\ &= g(x)h(y)\exp\left[\frac{\rho}{1-\rho^2}xy\right] = \alpha_i \gamma_j e^{\beta u_i v_j} . \end{aligned}$$

By the constraints of the RC model, $\alpha_i = \exp(\mu + \lambda_i^X)$, $\gamma_j = \exp(\lambda_j^Y)$ and the association parameter β correspond to $\rho/(1-\rho^2)$ in the bivariate normal density (see details in Goodman (1981b)).

If one set of the parameter scores are fixed, the RC model simplifies to the "row effects model" (R model) or "column effects model" (C model), where R model means the column scores v_j are fixed constants, and C model means the row scores u_i are fixed constants. Goodman (1979) suggested a simple iterative algorithm to fit the RC model using the method for the R model and C model. A cycle of the algorithm has two steps, each step consisting of the iterative fitting of the R or C model. First, the column parameter scores are treated as fixed values and estimate the row scores as in the R model. By using the estimated row scores from the first step, the column scores are estimated as in the C model. Second, using the estimated column scores, row scores are re-estimated. In each step, convergence is checked. When the criteria of convergence are satisfied, then the iterations are finished. Davis (1988) modified Goodman's algorithm. Therefore, Davis's algorithm is used to apply a bootstrap method to an $r \times c$ contingency table (see details in section 3 of Davis (1988)).

3. Bootstrap procedure for RC model

In this section, we propose the three bootstrap procedure to test for RC model. When the RC model holds, we focus to (1) testing of goodness of fit, (2) testing of $\beta = 0$, (3) testing whether row and column scores are equally distanced. Denote the "linear-by-linear model" by LL model and the "independence model" by I model.

Assuming the RC model holds, the test statistics of goodness of fit have the form $G^2(RC) = \sum\sum n_{ij}(\log n_{ij} - \log \hat{m}_{ij})$. The test has $(r-2)(c-2)$ degrees of freedom. Next, we assess the statistical significance of $H_0 : \beta = 0$ by testing the departure from independence. One test statistic is the reduction in G^2 statistic:

$$G^2(I|RC) = G^2(I) - G^2(RC).$$

Haberman (1981) showed that $G^2(I|RC)$ is not chi-squared, but instead the statistic has the same asymptotic distribution as that of the maximum eigenvalue of the $(r-1)(r-1)$ central Wishart matrix with $(c-1)$ degrees of freedom. Finally, testing whether the row and column scores are equally distanced, which is equivalent to testing for the departure from the "linear-by-linear model", we therefore use the statistic $G^2(LL|RC) = G^2(LL) - G^2(RC)$, based on $r+c-4$ degrees of freedom. Now, we propose a bootstrap test procedure to test for each case in the RC model.

A. Bootstrap procedure for goodness of fit test for RC model

[Step 1] Fit the RC model to the observed data $\{n_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ and compute the expected cell frequencies $\hat{m} = \{\hat{m}_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ by the Davis algorithm and expected cell probabilities $\hat{p} = \{\hat{p}_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$, where $\hat{p}_{ij} = \hat{m}_{ij}/n_{ij}$. Then compute the likelihood ratio goodness of fit statistic $G^2(RC)$.

[Step 2] Generate a bootstrap sample $\{n^*_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ from multinomial distribution with cell probabilities $\hat{p} = \{\hat{p}_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ based on the RC model.

[Step 3] Fit the RC model to the bootstrap sample $\{n^*_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ and compute the likelihood ratio goodness of fit statistic $G^{2*}(RC)$.

[Step 4] If B bootstrap replicates have been obtained, go to Step 5. Otherwise, repeat Step 2 and Step 3.

[Step 5] Compute the estimate of the p-value such as

$$\alpha_{Boot} = \# \ [G^{2*}(RC) \geqq G^2(RC)]/B.$$

B. Bootstrap procedure for testing $H_o: \beta = 0$ for RC model

[Step 1] Fit the RC model and the I model to the observed data $\{n_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$. Compute the expected cell frequencies $\hat{m} = \{\hat{m}_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ and the expected cell probabilities $\hat{p} = \{\hat{p}_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ for the RC model. Again compute the expected cell probabilities $g = \{g_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ for the I model where $g_{ij} = n_{i.}n_{.j}/n^2$. Using these two models, compute the testing statistic $G^2(I|RC) = G^2(I) - G^2(RC)$.

[Step 2] Generate a parametric bootstrap sample $\{n^*_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ from multinomial distribution with cell probabilities $g = \{g_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ based on the I model, instead of $\hat{p} = \{\hat{p}_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ based on the RC model.

[Step 3] Fit the RC model and the I model to the bootstrap sample $\{n^*_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$. Then compute the testing statistic $G^{2*}(I|RC) = G^{2*}(I) - G^{2*}(RC)$.

[Step 4] If B bootstrap replicates have been obtained, go to Step 5. Otherwise, repeat Step 2 and Step 3.

[Step 5] Compute the estimate of the p-value such as

$$\alpha_{Boot} = \# \ [G^{2*}(I) - G^{2*}(RC) \geqq G^2(I) - G^2(RC)]/B.$$

C. Bootstrap procedure for testing equality of intervals for RC model

[Step 1] Fit the RC model and the LL model to the observed data $\{n_{ij}: 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$.

Compute the expected cell frequencies $\widehat{m} = \{\widehat{m}_{ij} : 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ and the expected cell probabilities $\hat{p} = \{\hat{p}_{ij} : 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ for the RC model. Then compute the expected cell probabilities $\widehat{p_L} = \{\hat{p}^L_{ij} : 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ for the LL model. Using these two models, compute the testing statistic $G^2(LL|RC) = G^2(LL) - G^2(RC)$.

[Step 2] Generate a parametric bootstrap sample $\{n^*_{ij} : 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ from multinomial distribution with cell probabilities $\widehat{p_L} = \{\hat{p}^L_{ij} : 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ based on the LL model.

[Step 3] Fit the RC model and the LL model to the bootstrap sample $\{n^*_{ij} : 1 \leqq i \leqq r, 1 \leqq j \leqq c\}$ respectively. Then compute the testing statistic $G^{2*}(LL|RC) = G^{2*}(LL) - G^{2*}(RC)$.

[Step 4] If B bootstrap replicates have been obtained, go to Step 5. Otherwise, repeat Step 2 and Step 3.

[Step 5] Compute the estimate of the p-value such as

$$\alpha_{Boot} = \#\ [G^{2*}(LL) - G^{2*}(RC) \geqq G^2(LL) - G^2(RC)]/B.$$

There are some differences in generating the bootstrap sample in each case. When one execute the goodness of fit test for the RC model, testing the $H_0 : \beta = 0$, and testing the equality of intervals, then the bootstrap sample must be generated from multinomial distribution with the cell probabilities based on the RC model, I model, and LL model respectively.

4. Numerical Example

We consider a contingency table from Srole, Langner, Michael, Opler, and Rennie (1962) that describes the relationship between an individual's mental health status (MHS;$r = 4$) and the socioeconomic status of his or her parents (SES;$c = 6$). The data are displayed in Table 1. This example was also examined by Gilula (1986) and Goodman (1979, 1986). Here, the data are analysed using the bootstrap methods proposed in this paper.

The null hypothesis of independence between SES and MHS is not accepted, with the likelihood ratio chi-square $G^2(I) = 47.42$ based on $df = 15$. But the RC model fits the observed data very well, with $G^2(RC) = 3.57$ based on $df = 8$. The ML estimates of column scores are (-1.1124, -1.1215, -0.3711, 0.0271, 1.0104, 1.8181), the ML estimates of row scores are (-1.6773, -0.1405, 0.1367, 1.4139) and $\hat{\beta} = 0.1665$. The corresponding correlation parameter is $\hat{\rho} = 0.1622$. The row scores indicate that the distance between the MHS categories labeled "mild" and "moderate" is much less than the distances between other adjacent categories.

Also, SES categories A and B have almost the same score, and the distance between categories C and D is much less than the distances between B and C, D and E, and E and F.

[Table 1] Subjects Cross-Classified by Mental Health Status and Parental Socioeconomic Status

Mental Health Status (MHS)	Parental Socioeconomic Status (SES)						
	A(high)	B	C	D	E	F(low)	Total
Well	64	57	57	72	36	21	307
Mild Symptoms	94	94	105	141	97	71	602
Mod. Symptoms	58	54	65	77	54	54	362
Impaired	46	40	60	94	78	71	389
Total	262	245	287	384	265	217	1660

Now the bootstrap statistical inferences are considered for the scores of row and column categories, goodness of fit test for RC model, the estimated value of association parameter β and the test of equality of intervals of row and column scores.

First, the estimate of the bootstrap p-value of the goodness of fit is

$$P[G^{2*}(RC) \geqq G^2(RC)] = 0.894.$$

In comparison, the corresponding chi-squared p-value with $df = 8$ is 0.8937. Thus, the two methods produce quite similar results (based on bootstrap replication B=1000).

Second, the estimate of the bootstrap p-value of the hypothesis $H_0 : \beta = 0$ is

$$P[G^{2*}(I) - G^{2*}(RC) \geqq G^2(I) - G^2(RC)] = 0.000.$$

In comparison, $G^2(I) - G^2(RC) = 43.86$. The corresponding upper 5 and 1 percent points of the distribution of the maximum eigenvalue of the 3×3 central Wishart matrix with $df = 5$ are 17.21 and 21.65. Therefore, the p-value is preconceived as smaller than 0.01. Thus, the two methods produce quite similar results (based on bootstrap replication B=1000).

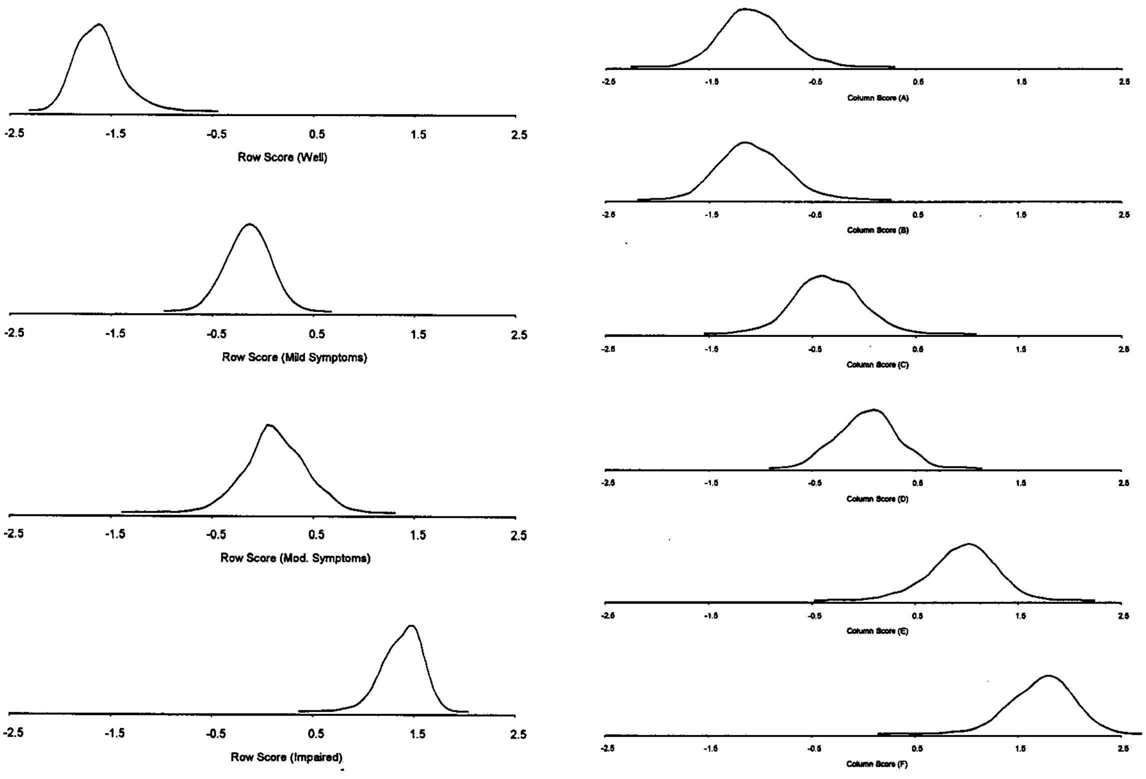
Third, the estimate of the bootstrap p-value of the hypothesis H_0: row and column scores are equally distanced, is

$$P[G^{2*}(LL) - G^{2*}(RC) \geqq G^2(LL) - G^2(RC)] = 0.391.$$

In comparison, when we use the statistic $G^2(LL) - G^2(RC) = 6.32$ based on $df = 6$, the p-value is 0.3883. This means that the parameter scores do not give a significantly better fit than equal-interval scores. However, in comparison with the bootstrap method, the two methods produce quite similar results (based on bootstrap replication B=1000).

[Table 1] was analyzed as it has been shown to have a stochastically ordering property by Evans et al. (1997). They analyzed the data with a Bayesian method. Hence, the bootstrap method produced the unexpected additional results which are the bootstrap distributions of the row scores, column scores, β and ρ values. The distributions can be expressed by the flexible and straightforwardly implemented feature of the bootstrap approach.

[Figure 1] displays the estimated kernel densities of row and column scores. [Figure 1(a)] goes from MHS categories labeled "Well" to MHS categories labeled "Impaired". [Figure 1(b)] goes from SES categories labeled "A" to SES categories labeled "F". The graphical outputs show the stochastically ordering property.

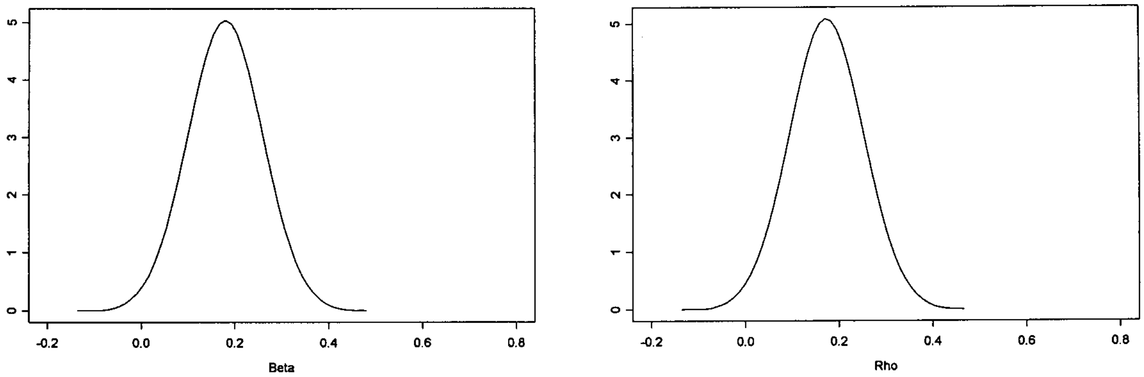


(a) Row scores

(b) Column scores

[Figure 1] Estimated kernel densities for row and column scores.

From [Figure 1(a)], we see that the densities of "Mild (rowscore 2)" and "Moderate (rowscore 3)" overlay in a wide range. This means there is only a slight difference between the row scores of "Mild" and "Moderate". In addition, from [Figure 1(b)], the densities of "A" and "B" show a similar pattern. Both [Figure 1(a)] and [Figure 1(b)] show that the distributions of scores are stochastically higher as row categories move from "Well" to "Impaired" and column categories move from "A" to "B". These graphic results can be applied to collapse the categories. Therefore, the "Mild" category can be combined with the "Moderate" category for the MHS variable. For the SES variable, categories "A" and "B" also can be combined. [Figure 2] displays the estimated kernel densities of β and ρ values. This two estimated densities show a similar pattern, and the values are distributed from the ML estimate 0.1665 and 0.1622 respectively.



(a) β values (b) ρ values

[Figure 2] Estimated kernel densities for β and ρ values.

5. Conclusion

Many difficulties lies in making statistical inferences because the RC model has a nonlinear property for parameters. In this study, we have proposed various bootstrap methods for ordered categorical data, especially for the RC model. The bootstrap methods is statistically more appealing because the complicated structures among the row and column categories are taken into account. We conclude that the proposed bootstrap methods provide a useful nonparametric alternative to the traditional asymptotical theory. Our approach can be generalized to the traditional log-linear model. Of course, we may extend these graphical results to the stochastic ordering among a set of categorical variables.

References

[1] Agresti, A. (1984) *Analysis of Ordinal Categorical Data*. New York : John Wiley & Sons.

[2] Agresti, A. (1990) Categorical Data Analysis. New York : John Wiley & Sons.

[3] Davis, C. S. (1988) Estimation of row and column scores in the linear-by-linear association model for two-way ordinal contingency tables. *Proceeding 13th Annual SAS Users Group International Conference*, 946-951.

[4] Evans, M., Gilula, Z., Guttman, I., and Swartz, T. (1997) Bayesian analysis of stochastically ordered distributions of categorical variables. *Journal of the American Statistical Association*, 92, 208-214.

[5] Gilula, Z. (1986) Grouping and association in two-way contingency table: A canonical correlation analytic approach. *Journal of the American Statistical Association*, 81,

773-779.

[6] Gilula, Z. and Ritov, Y. (1990) Inferential ordinal correspondence analysis : motivation, derivation and limitations. *International Statistical Review*, 58, 99-108.

[7] 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.

[8] Goodman, L. A. (1981a) Association models and canonical correlation in the analysis of cross classifications having ordered categories. *Journal of the American Statistical Association*, 76, 320-334.

[9] Goodman, L. A. (1981b) Associations models and the bivariate normal for contingency tables with ordered categories. *Biometrika*, 68, 347-355.

[10] Goodman, L. A. (1986) Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables. *International Statistical Review*, 54, 243-309.

[11] Haberman, S. J. (1981) Tests for independence in two-way contingency tables based on canonical correlations and on linear-by-linear interaction. *The Annals of Statistics*, 9, 1178-1186.

[12] Jeong, H.C., Jhun, M., and Kim D. (2005) Bootstrap tests for independence in two-way ordinal contingency tables, *Computational statistics & Data Analysis*, 48, 623-631.

[13] Srole, L., Langner, T. S., Michael, S. T., Opler, M. K., and Rennie, T. A. C. (1962). *Mental Health in the Metropolis : The Midtown Manhattan Study*, New York: McGraw-Hill.

[Received February 2005, Accepted July 2005]