The Korean Communications
in Statistics Vol. 3, No. 3, 1996
pp. 31-37

A Model Comparison Method for Hierarchical Loglinear Models

Hyun Jip Choi[1], Chong Sun Hong[2]

Abstract

A hierarchical loglinear model comparison method is developed which is based on the well known partitioned likelihood ratio statistics. For any pair of hierarchical loglinear models, we can regard the difference of the goodness of fit statistics as the variation explained by a full model, and develop a partial test to compare a full model with a reduced model in that hierarchy. Note that this has similar arguments as that of the regression analysis.

1. Introduction

Consider the hierarchical loglinear model to describe the data structures. It is hard to choose the well fitted loglinear model among all possible models. Consequently, many loglinear model selection methods have been suggested. One of the commonly used methods is to use the partitioned likelihood ratio statistics. We studied the properties of the partitioned statistics and their usage in many categorical data analysis textbooks (e.g. Harberman (1974), Bishop, Fienberg and Holland (1975), Fienberg (1983), Christensen (1990), Andersen (1991), Agresti (1990) and many others).

In this article, we develop a model comparison method by using analysis of variations based on the partition of the likelihood ratio statistics. For a multidimensional categorical data, the partitioned likelihood ratio statistics can be summarized into some tables which have the exactly same appearance as the ANOVA tables. Such suggested tables provide useful informations to explain the data structure and evaluate the effect of each term in the loglinear model. With these tables, several partial test statistics are proposed to test corresponding paired hierarchical loglinear models : one of each pair can be regarded as a full model and the other is a reduced model. Moreover, these tests are compared with the coefficients of determination of Christensen (1990) as a measure of the goodness of fit for a model.

1) Lecturer, Department of Applied Statistics, Kyonggi University, Paldal Gu, Suwon, 440-760, Korea
2) Associate Professor, Department of Statistics, Sung Kyun Kwan University, Seoul, 110-745, Korea

2. Partial tests for hierarchical loglinear model comparisons

In the regression analysis, the total sum of squares (SST) is defined as the sum of squares of deviations from the intercept of the regression line, i.e., the mean of the dependent variable. We may partition this sum of squares into two parts : one gives information on error which is referred to as the sum of squares of residuals (SSE), and the other does on the regression line referred to as the regression sum of squares (SSR). It is convenient to tabulate each component of the partition into a summary table called an analysis of variation (ANOVA) table, and the well known ANOVA table is used to evaluate the goodness of fit for the pre-designed model.

For categorical data, Gini (1912) had defined the total variation. Keeping with analysis of variation terminology, Light and Margolin (1971) partitioned Gini's total variation (TSS) into the within-group sum of square (WSS) and the between-group sum of squares (BSS) for the two-dimensional contingency table, and they defined the ratio BSS/TSS as a sample measure of the proportion of variation in the row variable which is attributable to the column variable.

Now suppose throughout this paper that our attention is restricted to the hierarchical loglinear models for complete contingency tables. We apply such partition method to hierarchical loglinear models and develop a partial test for multidimensional categorical data. The likelihood ratio statistic to test the goodness of fit for an assigned loglinear model,

$$G^2 = 2\sum x \log(x/\hat{m}) \tag{1}$$

is the variation between the maximum likelihood estimates ($\hat{m}$) and the observation values (x). We could define $G^2(a)$ as the likelihood ratio test statistic for loglinear model (a), and $G^2(b)$ is that of a hierarchical model (b) which includes model (a) with degrees of freedom d_1, and d_2, respectively. Under this hierarchy, it satisfies that $G^2(a) \geq G^2(b)$, and we can partition $G^2(a)$ into the following :

$$\begin{aligned} G^2(a) &= [G^2(a) - G^2(b)] + G^2(b) \\ &= 2\sum x \log(\hat{m}^{(b)}/\hat{m}^{(a)}) + 2\sum x \log(x/\hat{m}^{(b)}), \end{aligned} \tag{2}$$

where $\hat{m}^{(a)}$ and $\hat{m}^{(b)}$ are the MLEs for model (a) and (b), respectively.

From this partition, $G^2(b)$ can be regarded as the variation of errors which the estimates of model (b) do not explain the observed values, in other words, the unexplained variation by model (b). $G^2(a)-G^2(b)$ represents the improved goodness of fit by the effect of the terms added to model (b) from model (a), that is, the explained variation by model (b) in contrast with model (a). If we focus to explain the observed values through two hierarchcal log-linear models (a) and (b) satisfying $G^2(a) \geq G^2(b)$, then the variation $G^2(a)$ about the deviations by the smaller model (a) could be regarded as the total variation. Let us denote $G^2(b) \equiv SSE(b)$, $G^2(a)-G^2(b) \equiv SSR(a|b)$, and $G^2(a) \equiv SST(a)$. Then equation (2) can be rewritten as

$$SST(a) = SSR(a|b) + SSE(b). \tag{3}$$

For the above two hierarchical models (a) and (b), model (a) can be regarded as the reduced model which is smaller than model (b), and model (b) is the full model. In the analysis of the categorical data on two dimensional contingency table, the hierarchical structure only contains two models ; one is the complete independence model and the other is the full model which is the saturated model, so that $G^2(b) = 0$. Therefore, such partitioning technique in (2) will be meaningful for more than three dimensional categorical data.

Let us consider a pair of hierarchical loglinear models for more than three dimensional categorical data, and develop an analysis of variation based on the partition of the likelihood ratio statistic G^2 to evaluate the following models.

model (a) : a reduced model ,
model (b) : a full model which includes model (a).

According to the partition technique discussed in (2), we can summarize the partitioned terms into the following ANOVA table.

<Table 1> The ANOVA Table

S.V.	d.f.	Variation
$SSR(a\|b)$	d_1-d_2	$G^2(a)-G^2(b)$
$SSE(b)$	d_2	$G^2(b)$
$SST(a)$	d_1	$G^2(a)$

We note that $G^2(a)-G^2(b)$ and $G^2(b)$ in <Table 1> follow asymptotic χ^2 distributions with degrees of freedom d_1-d_2 and d_2, respectively. Hence the ratio of these two statistics

$$F^c = \frac{[G^2(a)-G^2(b)]/(d_1-d_2)}{G^2(b)/d_2} \tag{4}$$

follows an asymptotic F distribution with degrees of freedom d_1-d_2 and d_2. Therefore we can test the following hypothesis by using the F^c statistics.

H_0 : the reduced model (a)
H_1 : the full model (b)

The variation explained by model (b) in contrast with model (a), $SSR(a|b) = G^2(a)-G^2(b)$, might be due to the effect of the added terms from model (a) to model (b). If the effect of these added terms is significant for describing such a model structure, it indicates that there exists a big difference between goodness of fit of model (a) and model (b). Hence the large value of F^c statistic will guide us that the null hypothesis has to be rejected.

Now, we consider model (c) such that $G^2(b) \geq G^2(c)$ with the degrees of freedom d_3. Under this hierarchy, we can get the equations :

$$\begin{aligned} G^2(a) &= [G^2(a)-G^2(b)]+G^2(b) \\ G^2(b) &= [G^2(b)-G^2(c)]+G^2(c) \end{aligned}, \tag{5}$$

These equations can be summarized into the sequential ANOVA table which has the expanded form of <Table 1>.

<Table 2> The Sequential ANOVA Table

S.V.	d.f.	Variation
$SSR(a\|b)$	d_1-d_2	$G^2(a)-G^2(b)$
$SSE(b)$	d_2	$G^2(b)$
$SSR(b\|c)$	d_2-d_3	$G^2(b)-G^2(c)$
$SSE(c)$	d_3	$G^2(c)$
$SST(a)$	d_1	$G^2(a)$

In <Table 2>, $SSR(b|c)$ is the improved fit by model (c) in contrast with model (b). At this moment, we can regard model (b) as the reduced model and model (c) as the full model. Then, for model (b) and (c), the statistic in (4) is defined as

$$F^c = \frac{[G^2(b)-G^2(c)]/(d_2-d_3)}{G^2(c)/d_3} . \tag{6}$$

Also this statistic follows an aymptotic F distribution with degrees of freedom d_2-d_3 and d_3, so that we can test the following hypothesis.

H_0 : the reduced model (b)

H_1 : the full model (c)

Note that we may consider $SSR(a|b)$ and $SSR(b|c)$ as the type I sum of squares (sequential partial sum of squares) in the analysis of regression, so that we might call this method as the partial test in the analysis of loglinear model.

Christensen (1990) defined the coefficient of determination, R^2, as a measure of evaluating the goodness of fit for a hierarchical model (b) when model (a) is the smallest :

$$R^2 = \frac{G^2(a)-G^2(b)}{G^2(a)} = \frac{SSR(a|b)}{SST(a)} \equiv R^2_{(b)}. \tag{7}$$

$G^2(a)$ in the denominator of (7) could be regarded as the measure of the total variability in the data and $G^2(a)-G^2(b)$ in the numerator measures the variability explained by model (b). So R^2 of Christensen could be denoted as $R^2_{(b)}$ which is the proportion of the total variability explained by model (b). We can expand this fact for model (b) and model (c) and get

$$R^2_{(b|c)} \equiv \frac{G^2(b)-G^2(c)}{G^2(a)} . \tag{8}$$

Since the numerator is the explained variation by model (c) in contrast with (b), we will

call this $R^2_{(a|c)}$ as the partial coefficient of determination (see Choi and Hong (1995) for more details). If the F^c statistic in (4) or (6) has a large value, then $R^2_{(b)}$ or $R^2_{(a|c)}$ also has a considerable value.

In order to apply the sequential F^c testing method discussed previously, we take the well known $3\times2\times2\times2$ detergent preference data of Ries and Smith (1963). Through the combinations of the four categorical variables of the data, we can consider many hierarchical models. Since Bishop et. al. (1975), Fienberg (1983), and others considered the following six hierarchical loglinear models for model selection, we consider the same hierarchy.

<Table 3> Goodness of fits for Detergent Preference data

ID	MODEL	d.f.	G^2	Differences	d.f.	G^2
(a)	[1][2][3][4]	18	42.93*			
(b)	[1][3][24]	17	22.35	(a) and (b)	1	20.58*
(c)	[1][24][34]	16	17.99	(b) and (c)	1	4.36*
(d)	[13][24][34]	14	11.89	(c) and (d)	2	6.10*
(e)	[1][234]	12	8.41	(d) and (e)	2	3.48
(f)	[123][234]	8	5.66	(e) and (f)	4	2.75

* indicates that the p-value of the statistic is less than 5% significant level.

From the result in the right side of <Table 2>, one might choose all models which include model (b) as the well fitted model . However, the left side of <Table 3> suggests that the model (d) is the best in the given hierarchy. Now we can apply the comparison test for the adjacent pair of models sequentially and the results are summarized into <Table 4>.

<Table 4> The Sequential ANOVA Table

Comparison	F^c	p-value
with model (a) and (b)	15.65	0.001
with model (b) and (c)	3.88	0.066
with model (c) and (d)	3.59	0.055
with model (d) and (e)	2.48	0.125
with model (e) and (f)	1.01	0.458

As a result in <Table 4>, model (b) would be selected to the well fitted model at 5% significant level. However, if we set up the significant level as 10%, the same result can be obtained as that of <Table 3>.

3. Concluding remark

We developed a test for comparing a reduced and a full loglinear models in a given hierarchical structure. Since the suggested test statistic is based on the ratio of an explained and an unexplained variations, it can determined the effect of the added terms to the full model for describing the data structure. Furthermore, we may adapt the sequential tests to model selection method for hierarchical structures. We can possibly say that this test statistics is more conservative than the partitioned statatistic which is commonly used for model selection. Nonetheless the conservativeness is not a crucial defect for the comparisons since the selection processes have involved many decision strategies and a well fitted model for a given data may differ from personal opinions.

References

[1] Agresti, A. (1990). *Categorical Data Analysis*, John Wiley & Sons.

[2] Andersen, E. B. (1991). *The Statistical Analysis of Categorical Data*, Springer-Verlag.

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

[4] Choi, H. J. and Hong, C. S. (1995). Graphical descriptions for hierarchical loglinear Models, *The Korean Communications in Statistics*, Vol. 2, No. 2, 310-319.

[5] Christensen, R. (1990). *Log-Linear Model*, Springer-Velag.

[6] Fienberg, S. E. (1983). *The Analysis of Cross-Classified Categorical Data*, MIT Press.

[7] Gini, C. (1912). Variabilit e mutabilit, contributo allo studio delle distribuzioni ; relazione statische, In Studi Economico-Giuridici della R. Universit di Carliari.

[8] Harberman, S. J. (1974). *The Analysis of Frequency Data*, The University of Chicago Press.

[9] Light, R. J. and Margolin, B. H. (1971). An analysis of variance for categorical data, *Journal of American Statistical Association*, 66, 534-544.

[10] Ries, P. N. and Smith, H. (1963). The use of chi-square for preference testing in multidimensional problems, *Chemical Engineering Progress*, 59, 39-43.