Jour. Soc. Sci. & Humn. 1994/95: 1 & 2, 53-60

LOG-LINEAR AND LOG-BILINEAR MODELS WITH ORDINAL CATEGORIES

JUNAID S. SIDDIQI
Department of Statistics
University of Karachi, Karachi-75270, Pakistan

Log-linear models provide a natural analog to the classical least square regression for categorical frequency data. These models are also extended for the data having ordered category. In this article ordinal log-linear models and RC-association (log- bilinear) model are compared. Also presented an idea for the interpretation of the parameters on the basis of odds-ratio. These models were applied to a data to illustrate the computation and comparability.

1. Introduction:

There are several techniques in multivariate methods that can be used for the purpose of data reduction and display of the multivariate data in a lower dimensional space; for example principal component analysis, multi-dimensional scaling, and reciprocal averaging. These methods deal with the cases with metrical observation while correspondence analysis and log-linear models can be used for categorical data. In this paper the aim is to study the log-linear model in special context of ordinal categorical data. For details on other methods see Mardia et al (1979), Gifi (1981), Greenacre (1984), Anderson (1988) and Krzanowski (1989).

The idea of log-linear models is to provide a natural analog to the classical least square regression for the categorical frequency data. In the least square model the symmetric effects are assumed to be additive, while in log-linear model they are assumed to have multiplicative effect. The normality assumption is replaced by the multinomial or Poisson. Like the classicial regression model, the purpose of the log-linear model is to describe the association among the categorical variables. For the usual log-linear models Haberman (1974), Duncan (1979), Duncan and Rao (1979), Agresti (1990) and Haggaman (1991) can be referred. In this article we present a brief survey of the ordinal log-linear model and its interpretation for two variables X and Y where X = 1, 2,... i,..., I and Y = 1, 2, ..., j ,..., J. In the following sections we consider three types of ordinal models, namely one with either X

category or Y category ordinal or both X and Y categories ordinal the third model also consider both X and Y categories ordinal but the order of the levels are not fixed by the investigator but treat them as parameters.

2. Log-linear Models for Ordinal Variables:

The classical log-linear model treats all the variables in the cross-classification as nominal, both for the purpose of estimation and goodness of fit. This limitation is serious because ordinal data is very common and appropriate models should obviously reflect the ordering. Ordinal categorical data frequently arise in every field of applied research, and may reflect categorical response which has an ordered set of levels or else arise from the grouping of continuous response such as age. Variables with ordinal categories often arise in social research, for example, in 1975 General Research Survey (Davis, 1975) almost half of the 230 recorded variables were ordinal. In many surveys the responses are typically of the following type:

(i) (a) Very happy (b) Pretty happy (c) Not happy

(ii) (a) Far below average (b) Below average (c) Average (d) Above average (e) Far above average

(iii) (a) A very great deal (b) A great deal (c) Quite a bit (d) A fair amount (e) Some (f) A little (g) None

Clayton (1974) proposed some methods for two way data based on odds ratio and his idea was generalized for multi-way table by serveral authors, such as McCullagh (1974, 1978), Anderson and Philips (1981) Agresti (1990) and Siddiqui (1992). However a log- linear model for ordinal data can be obtained by imposing suitable restriction on the effects parameters. A specialized model for the analysis of assoication in cross-classification table with ordinal categories has been developed by Goodman (1979) usually referred as row-column (RC) association model.

These model study the effects of row and column categories on the expected frequencies in the cross-classification table dentoted by mij,

2.1 The Ordinal-Nominal Model:

Suppose that the row variable X is ordinal and the column variable is nominal. Let $\{u_i\}$; $i = 1, ..., I$ denote a set of assigned row scores and they are subjected to the following restriction

$$u_1 \leq ... \leq u_i - ... u_I \qquad (1)$$

A possible ordenal model is:

$$\log [m_{ij}] = u + \lambda_i^X + \lambda_j^Y + \tau^Y (u_i - \bar{u}) \qquad (2)$$

where u 's are assumed to be equidistant and known and the parameters hold the restrictions that

$$\sum_{i=1}^{I} \lambda_i = \sum_{j=1}^{I} \lambda_j = \sum_{j=1}^{J} \tau_j = 0$$

Now the model is completely identifiable and unsaturated and has (I-1) + 2 (J-1) + 1 parameters. The parameters λ^X and λ^Y represent the marginal effects of X and Y on the expected frequencies respectively. The association between X and Y is explained by π^Y and this reflects the ordinal nature of X.

The ordinal nature of X can be seen and interpreted using the idea of the log-odds ratios. The log-odds between a pair of rows i' and i (say i > i) is:

$$\log \left[\frac{m'_{ij}}{m_{ij}} \right] = \tau_{j.}^{Y} (u_i - u_{i'}) \qquad (3)$$

This indicates that for $\tau_{j.}^{Y} > 0$ the probability of having large frequencies in the succeeding rows is higher in the j-th column. The log-odds ratio between pairs of rows i and i and columns j and j' ($j > j'$), is

$$\log \left[\frac{m_{ij} / m_{i'j}}{m_{ij'} / m_{i'j'}} \right] = (\tau_{j.}^{Y} - \tau_{j'.}^{Y}) (u_j - u_{j'}) \qquad (4)$$

That is the log-odds ratio is proportional to the distance between the two rows and if $(t_{j.}^{Y} - t_{j.}^{Y},) > 0$ this say there is a higher chance of having large frequencies towards the lower side j-th column than the j-th column.

The model for the ordinal column can be expressed as follows

$$\log (m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \tau_i^X (v_j - v) \qquad (5)$$

This model can be interpreted in a similar fashion as we did in the case of the model (2).

2.2 The Ordinal-Ordinal Model:

We now assume that both the row and column categories are ordinal and that $\{u_i\}$ and $\{v_j\}$ are ordered row and column score with

$$u_1 \leq \ldots \leq u_i \leq u_I \text{ and } v_1 \leq \ldots \leq v_j \leq \ldots \leq v_J \qquad (6)$$

Then a log-linear ordinal model can be written as

$$\log (mij) = \mu + \lambda_i^X + \lambda_j^Y + \beta_x^Y (u_i - \bar{u}) (v_j - \bar{v}) \qquad (7)$$

This model has only one more parameter than the idnependence model so the degress of freedom are (I-1) + (J-1) + 2. The log- odds ratio between the rows and columns as described above is now

$$\log \left[\frac{m_{ij}/m_{i'j}}{m_{ij'}/m_{i'j'}} \right] = \beta^{xy} (u_i - u_{i'}) (v_j - v_{j'}) \qquad (8)$$

where the parameter β_x^Y reflects the linear dependence of the log-odds ratio on both the distance between row and column scores. This model decomposes the departure form the independence, $\beta_x^Y = 0$ indicated that the variables are independent.

2.3 RC Association Model:

The model we have discussed in sections 2.1 and 2.2 are useful when either the row or column categories or both are ordinal and the categories can be assigned, usually equidistant, scores by the reseracher. Use of different column or row score will generally give completelly different estimates of the parameters and possibly differenct conclusion on the strength of the association. In particular this will happen when scores corresponding to a different ordering of the categories are used. If the data do indeed follow the ordering of the researcher's choice of scores the model's performance will be better than that under other implied orderings. This idea was used by Goodman in developing an alternative model under the assumption that both row and column categories are ordinal but the ordering is unknown, see Goodman (1979a, 1981a and 1981b). This model treats the row and column categories as unknown parameters to be estimated from the data and is invariant to the labelling of the categories of the rows an column of the table. The RC association model can be expressed as

$$\log (m_{ij}) = \mu + \lambda_i^X + \lambda_{jy} + \gamma_x^Y \mu_i \upsilon_j \qquad (9)$$

where λ's are the main effects parmeters as in models (2) and (7) and γ measures the interaction effects between X and Y. The row and column scors μ's υ's are normalized as follows

$$\sum_{i=1}^{I} \mu_i = \sum_{j=1}^{J} u_j = 0 \qquad (10)$$

$$\sum_{i=1}^{I} \mu_i^2 = \sum_{j=1}^{J} v_j^2 = 1 \qquad (11)$$

Because of treating the row and column scores as parameters the degrees of freedom reduce to (I-2)(J-2). This model can be interpreted in the same way as model (7), except that here the row and column scores may provide some additional information on the ordering of the categories.

The log-odds for adjacent rows i and i + 1 in the j-th column is

$$\log [m_{ij} / m_{i\,j+1}] = \gamma^{XY} v_j (\mu_i - \mu_{i+1}) \qquad (12)$$

The γ^{XY} is proportional to the log-odds, but the interpretation is not as simple as in the ordinal-ordianl model, because there is now no restriction on the μ_i s. Similarly the log-odds for adjacent column j and j + 1 is

$$\log [\, m_{ij} / m_{i\,j+1}] = \gamma^{XY} \mu_i (v_j - v_{j+1}) \qquad (13)$$

Therefore the log-odds ratio between the adjacent rows and columns is

$$\log \left[\frac{m_{ij} / m_{i+1\,j}}{m_{ij} / m_{i\,j+1}} \right] = \gamma^{XY} (v_j - v_{j+2}) (\mu_i - \mu_{i+1}) \qquad (14)$$

The log-odds ratio will depend on the score and there fore in interpreting the model one also needs to account for the row and column scores. It is suggested that after fitting the RC- assocication model the row and column categories of the contingency table are re-arranged in ascending or descending order of the scores, although generally if the categories have a natural ordering the scores will reflect this. Once arranged in this way the scores are monotonic and the model may be interpreted in the same way as ordinal-ordinal model.

3. Example:

The data is based on a survey conducted by the Economic History department of the University of Exeter in an attempt to review economics and social history in British Universities. The actual questionnaire contained 51 questions but we selected 15 questions for the analysis. The question took the form of statement on which respondents were asked to give their opinion in terms of levels of agreement or disagreement. The opinion was measured on a five-point scale with levels being Agree Strongly (SA), Agree (AG), Neither Agree Nor Disagree (AD), Disagree (DA) and Strongly Disagree (SD). The

respondent were 174 university teachers of whom 86 belong to Economic History (EH) department, 32 to Economics (EC), 37 to History departments and the remaining 19 to some other (OT) departments. The responses of 174 teachers on the 15 selected question were cross-classified according to opinion (SA, AG, AD, DA and SD) and the teaching department (EC,EH,HI, and OT) which resulted the following 5 X 4 table

Table 1: Cross-tabulation of Department and Opinion

	Department			
Opinion	EC	EH	HI	OT
SA	44	133	59	14
AG	178	448	169	88
AD	108	257	88	65
DA	109	376	185	100
SD	25	71	31	09

The total number of the responses of 174 individual on the 15 questions should be 2610 but because of 57 missing responses the total is 2553.

To illustrate independence and all log-linear ordinal models were fitted to the above cross-classification table and the estimates of relevant parameters and deviances are presented in table 2.

Among these log-linear and bilinear models, the ordinal department model (Model 3) seems to be the best fit with a chi-square value ($x2_{(.05)}$ = 20.09) at same degrees of freedom. The model treating opinion as ordinal gives a relatively poor fit. The RC model (Model 5) has the minimum deviance but fewer degrees of freedom and so the fit is not better than department ordinal model. For the interpretation of the parameters on the basis of log odds can be done using the relevant equations.

Table 2: Estimation of log-linear and RC-association models

	Model Estimates	d.f.	Deviance	Mean dev.	$x^2(.05)$
1.	Independence	12	36.16	3.183	26.22
2.	Opinon ordinal $\tau_1^x = -.09898$ $\tau_2^x = -.03518$ $\tau_9^x = .02763$ $\tau_4^x = +.10653$	9	27.42	3.047	21.67
3.	Depmartment ordinal $\tau_1^y = -.06348$ $\tau_2^y = -.05128$ $\tau_3^y = -.00848$ $\tau_4^y = +.18072$ $\tau_5^y = -.05748$	8	17.26	2.158	20.09
4.	Both ordinal $\beta^{xy} = .0671$	11	27.46	2.490	23.21
5.	RC- assoication $\mu_1 = -.1698$ $\mu_2 = -.3079$ $\mu_3 = -.2207$ $\mu_4 = +.8895$ $\mu_5 = -.1912$	6 $v_1 = -.7441$ $v_2 = -.1534$ $v_3 = +.3489$ $v_4 = +.5486$	16.56	2.76	16.81

REFERENCES

Agresti,(1990). Categorical data analysis. N.Y. Academic Press. Anderson, T.W. (1988). An introduction to multivariate statistical analysis. N.Y. John Wiley.

Duncan (1975). Introducation to structural equation models. N.Y. Academic Press.

Gifi (1981). Non-linear mulitivaraite analysis. Leiden. Department of Datatheory.

Goodman, L.A. (1979). Simple models for the analysis of association in cross-classification having ordered category. J. Amer Statist., 74, 537-52.

Goodman, L.A. (1981a). Association models and canonical correlation in the analysis of cross-classification having ordered categories. J. Amer. Statist. Assoc. 76, 320-34.

Goodman, L.A. (1981b). Association models and the bivariate normal distribution in the analysis of cross-classification having ordinal categories. Biomerika, 68, 347-55.

Greenacre (1984). Theory and application of correspondence analysis. London. Academic Press.

Haberman, (1974). Log-linear models for frequency table derived by indirect observation: Maximum likelihood equations. Ann. Statist., 2, 911-22.

Haberman (1979). Analysis of qualitative data Vol. 3: New developments. N.Y. Academic Press.

Krzanowski, W. (1989). Principals of multivariate analysis: A user's perspective. Oxford Clarenden Press.

Mardia, Kent and Bibby (1979). Multivariate analysi. London. Academic Press.

Siddiqi, J. S. (1992). Mixture and latent class models for multivaraite discrete data. Ph.D. Thesis; University of Excter.