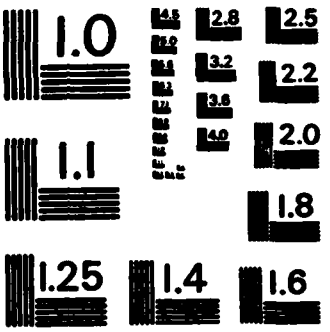
AD-A121 923 USING INFORMATION ON ORDERING FOR LOGLINEAR MODEL ANALYSIS OF MULTIDIMENS.. (U) CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF STATISTICS S E FIENBERG JUN 82 TR-255 N00014-80-C-0637 F/G 12/1 1/1

UNCLASSIFIED NL

END

FILMED

DTIC



(12)

USING INFORMATION ON ORDERING FOR LOGLINEAR MODEL ANALYSIS OF MULTIDIMENSIONAL CONTINGENCY TABLES*

by

Stephen E. Fienberg

DEPARTMENT

OF

STATISTICS

Carnegie-Mellon University

PITTSBURGH, PENNSYLVANIA 15213

12

USING INFORMATION ON ORDERING FOR LOGLINEAR MODEL ANALYSIS OF MULTIDIMENSIONAL CONTINGENCY TABLES*

by

Stephen E. Fienberg

Technical Report No. 255

Department of Statistics

Carnegie-Mellon University

Pittsburgh, PA 15213 USA

June, 1982

*TO APPEAR IN THE PROCEEDINGS OF THE XITH INTERNATIONAL BIOMETRICS CONFERENCE, TOULOUSE, FRANCE, SEPTEMBER 6-11, 1982. THE PREPARATION OF THIS PAPER WAS SUPPORTED IN PART BY THE OFFICE OF NAVAL RESEARCH UNDER CONTRACT N00014-80-C-0637.

SUMMARY

Many different authors have claimed that the loglinear model approach to the analysis of contingency table data is appropriate only for nominal variables and does not make use of information on the ordinal nature of some categorical variables (i.e. the ordering of the categories). In this paper, we reviews a variety of loglinear model methods which do take into account, either explicitly or implicitly, such information on ordering. The Our focus is on methods involving maximum likelihood estimation, but other methods of estimation can be used with these models. We also consider briefly, Some additional models for ordered categorical data. are considered.

KEY WORDS: Categorical data; Loglinear models; Maximum likelihood estimates; Multidimensional contingency tables; Ordered categories; Ordinal variables.

1. INTRODUCTION

Following important theoretical work by Birch, Bishop, Goodman and others in the 1960's, there has been a resurgence of interest in the analysis of categorical data, especially in the form of multidimensional cross-classifications or contingency tables. Much of this recent literature has focussed on the use of loglinear models for tables of expected values, and detailed descriptions of maximum likelihood estimation methods for loglinear model analysis can now be found in numerous books (e.g. see BISHOP, FIENBERG, and HOLLAND (1975) and FIENBERG (1980)). It has been implied or even explicitly suggested by many authors that the use of these loglinear models is inappropriate when one or more of the variables involved is ordinal, i.e. has categories which have an ordered structure. The purpose of this paper is to demonstrate that these criticisms of loglinear model approaches to the analysis of categorical data are false.

For example, McCULLAGH (1980) has claimed that a general property of all loglinear models that do not use scores (see section 2 for a description of such models) is that they are permutation invariant, i.e. that the categories of variables can be permuted in an arbitrary way without affecting the fit or values of the parameters. In fact, the ordinal nature of some categorical variables is often crucial to the structural organization of categorical data subjected to loglinear analysis, as in triangular arrays (e.g., see BISHOP and FIENBERG (1969) or BISHOP, FIENBERG, and HOLLAND (1975), Chapter 5), and social mobility tables (e.g., see social mobility tables GOODMAN (1972, 1979a) or BISHOP, FIENBERG, and HOLLAND (1975), Chapters 5 and 9), and age-period-cohort structures (FIENBERG and MASON (1978)). The fit of loglinear models to these and other structures is *not* permutation invariant.

For our discussion here we distinguish between response and explanatory variables (see FIENBERG (1980), Chapter 1), and describe in Sections 2 and 3 some simple and direct loglinear model approaches which explicitly take into account the ordinal structure of explanatory and response variables, respectively. Then we turn, in Section 4, to a special class of nonlinear extensions to loglinear models where the ordering of categories lends a simple interpretation of interaction terms.

Finally, in Section 5, we mention several other approaches to the analysis of ordinal data that are related in some ways to loglinear model methods.

In addition to the formal use of ordering in the loglinear models mentioned above, and described in later sections, we should not lose sight of the fact that information on ordering of categories can be used in informal ways as well (e.g. see FIENBERG (1980), Chapter 3, p.46).

2. SCORES FOR ORDINAL EXPLANATORY VARIABLES

SIMON (1974), HABERMAN (1974), and FIENBERG (1980) describe a loglinear model approach in the case where one or more variables are ordinal, and the categories of these variables have preassigned scores. We consider one such model here in case of a $2 \times J \times K$ table, with variable 1 being a binary response variable and variables 2 and 3 being explanatory variables with ordered categories, the scores for which are $\{v_j^{(2)}\}$ and $\{v_k^{(3)}\}$, respectively. We begin with the loglinear model

$$\log m_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}. \tag{2.1}$$

Next we assume that the two-factor effects relating the explanatory variables to the response variable reflect the ordering given by the following restrictions:

$$u_{12(ij)} = (v_j^{(2)} - \bar{v}^{(2)})u_{1(i)}^{(2)} \tag{2.2}$$

and

$$u_{13(ik)} = (v_k^{(3)} - \bar{v}^{(3)})u_{1(i)}^{(3)}. \tag{2.3}$$

The usual hierarchical restriction on "ordered effects" then implies that, for the 2nd-order interaction terms,

$$u_{123(ijk)} = (v_j^{(2)} - \bar{v}^{(2)})(v_k^{(3)} - \bar{v}^{(3)})u_{1(i)}^{(23)}. \tag{2.4}$$

The terms $u_{1(i)}$, $u_{1(i)}^{(2)}$, $u_{1(i)}^{(3)}$, and $u_{1(i)}^{(23)}$ may all be different.

If we substitute for expressions (2.2), (2.3), and (2.4) in (2.1), and reexpress the model in terms of logits, we get

$$\text{logit}_{jk} = \log \frac{m_{1jk}}{m_{2jk}} = b_0 + b_2 v_j^{(2)} + b_3 v_k^{(3)} + b_{23} v_j^{(2)} v_k^{(3)} . \qquad (2.5)$$

where $b_2 = 2u_{1(1)}^{(2)}$, $b_3 = 2u_{2(1)}^{(2)}$, $b_{23} = 2u_{1(i)}^{23}$, and $b_0 = 2u_{1(1)} - b_2\bar{v}^{(2)} - b_3\bar{v}^{(3)} - b_{23}\bar{v}^{(2)}\bar{v}^{(3)}$. Expression (2.5) is a logistic regression model with an interaction term involving the two sets of scores. The key to this model is the use of the scores $\{v_j^{(2)}\}$ and $\{v_k^{(3)}\}$ to induce metrics for the two explanatory variables.

The likelihood equations for model (2.5) follow the standard structure for loglinear models, and are found by setting sufficient statistics equal to their expected values. These equations can be solved using one of a number of standard numerical techniques.

The method suggested here for 2 × J × K tables can be extended to more explanatory variables, and can be used when some predictor variables are nominal and others are categorical. It can even be used in cases where the response variables are ordinal and have pre-specified scores associated with them.

When pre-specified scores are not available, it may make sense to attempt to estimate the scores for one or more variables so as to optimize some criterion function. NISHISATO (1980) gives a detailed description of one such approach known as dual scaling. This approach is known in France under the name *l'analyse des correspondances* (correspondence analysis) and has been developed by Benzecri and his associates (BENZÉCRI, 1973). Nothing in the description of this section requires that the pre-assigned scores have an ordering that goes with the ordering of the categories of the corresponding variable, although in practice this will almost always be the case. NISHISATO (1980, Chapter 8) describes a dual-scaling approach to ensure order restrictions on the scores, which involves a version of the "pool-adjacent-violators" algorithm used in many other applications (BARLOW, BARTHOLOMEW, BREMMER, and BRUNK, 1972).

3. CONTINUATION RATIOS FOR AN ORDINAL RESPONSE VARIABLE

Suppose we have a single I-category ordinal response variable and a pair of explanatory variables (with J and K categories respectively). The general loglinear model given by expression (2.1), can be rewritten as a set of I-1 simultaneous logit models, e.g. for log (m_{ijk}/m_{Ijk}) for i = 1,2,...I-1, but this model does not reflect the ordinal nature of the response variable. A natural alternative to (2.1) and the resulting simultaneous logit models is to focus on *continuation ratios* of the form

$$m_{ijk}/\Sigma_{h>i} m_{hjk} \quad i = 1,2,...,I-1, \tag{3.1}$$

or

$$m_{ijk}/\Sigma_{h<i} m_{hjk} \quad i = 2,3,...,I. \tag{3.2}$$

Logit-like models for these continuation ratios can then be modelled separately because the likelihood function is a product of I-1 components, one for each continuation ratio.

A set of I-1 logit models for the continuation ratios may treat the explanatory variables differently in each of the I-1 equations, and thus the "effects of the ordered structure" are allowed to come through. FIENBERG (1980) gives an example of this approach with three explanatory variables, and illustrates that modelling a data-set with continuation-ratios going in one direction, as in expression (3.1), will not necessarily yield the same results as modelling those going in the other direction, as in expression (3.2).

The continuation-ratio approach suggested in this Section can be combined with methods for treating explanatory variables as ordinal, such as those described in Section 2. For example, FIENBERG and MASON (1978) use the continuation-ratio approach to analyze educational attainment in the U.S. by developing logit models with simultaneous effects for age, period, and cohort.

Although the "combined model," merging together the I-1 logit models for the continuation ratios, is not itself loglinear, it can be fit by standard maximum likelihood methods for loglinear or logit models applied separately to each of the logit models. As McCULLAGH (1980) notes, this approach is especially well-suited to problems where the response variables are discrete, and where it is unwise to try to treat them as course groupings of some finer scale.

4. SOME NONLINEAR EXTENSIONS OF LOGLINEAR MODELS

Again let us consider a $2 \times J \times K$ table where variable 1 represents a binary response, and variables 2 and 3 are explanatory. If we begin with the logit model of expression (2.1), and assume that only the 2nd-order interaction term depends on the pre-assigned scores, as in (2.4), then we have the linear logit model

$$\log\left(\frac{m_{1jk}}{m_{2jk}}\right) = w + w_{2(j)} + w_{3(k)} + b_{23} v_j^{(2)} v_k^{(3)}. \tag{4.1}$$

If $v_j^{(2)}$ is proportional to j and $v_k^{(3)}$ is proportional to k, expression (4.1) is a logit generalization of what GOODMAN (1979b) refers to as the uniform association model. Next suppose we do not have pre-assigned scores, and thus wish to estimate $\{v_j^{(2)}\}$ and $\{v_k^{(3)}\}$. To ensure the identifiability of these new parameters in the model, we need two constraints such as

$$\Sigma_j (v_j^{(2)})^2 = \Sigma_k (v_k^{(3)})^2 = 0 \tag{4.2}$$

in addition to the usual ANOVA constraints. This model is a linear logit model generalization of a model suggested for use in two-way tables by FIENBERG (1968), and studied in detail by GOODMAN (1979b). When $b_{23} = 0$, (4.1) reduces to the usual no 2nd-order interaction model. When the $\{v_j^{(2)}\}$ are proportional to $\{w_{2(j)}\}$ and the $\{v_k^{(3)}\}$ are proportional to $\{w_{3(k)}\}$, then we can dispose of the added constraint (4.2), and we get a nonlinear logit version of Tukey's 1 degree of freedom model for nonadditivity:

$$\log\left(\frac{m_{1jk}}{m_{2jk}}\right) = w + w_{2(j)} + w_{3(j)} + \lambda\, w_{2(j)} w_{3(k)}. \tag{4.3}$$

CHUANG (1980) gives details on the maximum likelihood estimation of parameters for both (4.1) and (4.3).

Note that both models (4.1) and (4.3) are invariant under changes of row and column orderings. But, as AGRESTI (1982) notes, there is a simple interpretation of the model when the $\{v_j^{(2)}\}$ and $\{v_k^{(3)}\}$ in model (4.1) are monotonic. Suppose that $v_1^{(2)} < v_2^{(2)} < \ldots v_J^{(2)}$. Then, if $v_k^{(3)} > v_k^{(3)}$, the log-odds for being in category 1 of the response variable are always greater for those observations for

which variable 3 takes the value k, then those for which variable 3 takes the value k'. Agresti suggests that, when we expect such stochastic orderings as a result of the ordinal structure of variables, we should add the appropriate inequality constraints to the model, although he does not explain how to get maximum likelihood estimates in those cases. Once the ordering is required we lose most of the invariance in the model, and are left with only palindromic invariance (McCULLAGH, 1978), associated with completely reversing the categories of the ordered variables.

As in the case of the models of Sections 2 and 3, we can readily generalize the models to more than two explanatory variables, and to mixtures of nominal and ordinal explanatory variables. Moreover, we can use such models, when the response variable is polytomous but ordinal, for continuation ratios of the sort described in Section 3. Finally, we can use the same type of model structure for multiple, polytomous ordinal response variables, expressing the parameters of interest as functions of the effects of explanatory variables.

5. OTHER METHODS

In this review, we have focussed on models that either are loglinear in nature, or involve natural extensions to loglinear models. Many alternative approaches are available. McCULLAGH (1980), for example, develops models for ordinal response variables which replaces the log odds of either version of the continuation ratio (expressions (3.1) or (3.2)) by the "accumulated" logit

$$\log \left[\Sigma_{h \leq i} m_{hjk} \; / \; \Sigma_{h > i} m_{hjk} \right] . \qquad (5.1)$$

Then he models these accumulated logits using linear models. This approach typically yields a stochastic ordering of response variables, and the same model is then applicable when categories of the response variable are collapsed. Unfortunately, the accumulated logits cannot be analyzed independently, and thus McCullagh's approach is likely to lead to computation problems when the table being analyzed is large. As McCullagh notes, (5.1) can easily be generalized through the use of a "link" function other than the logit.

Yet another way to approach information on orderings is to incorporate it into the modelling in

the form of order restrictions. BARLOW, BARTHOLOMEW, BREMNER, AND BRUNK (1972) describe several results related to binomial and multinomial problems. Most of this and subsequent literature focusses on maximum likelihood estimation when there is a stochastic order restriction on the probabilities themselves. EDDY, FIENBERG, and MEYER (1982) have developed a new approach wherein the order restrictions implied by the ordinal categorical variables are placed on marginal totals of the cross-classification. There appears to be an interesting link between this approach and ideas associated with loglinear models.

The preparation of this paper was supported in part by the Office of Naval Research under Contract N00014-80-C-0637.

6. REFERENCES

AGRESTI, A., A survey of strategies for modelling cross-classifications having ordinal variables, Unpublished manuscript, 1982.

BARLOW, R.E., BARTHOLOMEW, D.J., BREMNER, J.M., and BRUNK, H.D., Statistical inference under order restrictions: The theory and application of isotonic·regression, Wiley, New York, 1972.

BENZÉCRI, J.-P., L'analyse des correspondances (Volume 2 of L'analyse des donnees), Dunod, Paris, 1973 (in French).

BISHOP, Y.M.M. and FIENBERG, S.E., Incomplete two-dimensional contingency tables, Biometrics, 25, 119-128, 1969.

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W., Discrete multivariate analysis: Theory and practice, MIT Press, Cambridge, Mass., 1975.

CHUANG, J.-L.C., Analysis of categorical data with ordered categories, Ph.D. Dissertation, School of Statistics, University of Minnesota, 1980.

EDDY, Wm. F., FIENBERG, S.E., and MEYER, M.M., Contingency table estimation with order restrictions on the margins, Unpublished manuscript, 1982.

FIENBERG, S.E., The estimation of cell probabilities in two-way contingency tables, Ph.D. Dissertation, Department of Statistics, Harvard University, 1968.

FIENBERG, S.E., The analysis of cross-classified categorical data (2nd edition), MIT Press, Cambridge, Mass., 1980.

FIENBERG, S.E. and MASON, W., Identification and estimation of age, period and cohort models in the analysis of discrete archival data, Sociological Methodology 1979, 1-67, 1978.

GOODMAN, L.A., Some multiplicative models for the analysis of cross-classified data, Proc. 6th Berkeley Symp. Math. Statist. Prob., 1, 649-696, 1972.

GOODMAN, L.A., Multiplicative models for square contingency tables with ordered categories, Biometrika, 66, 413-418, 1979a.

GOODMAN, L.A., Simple models for the analysis of association in cross-classifications having ordered categories, J. Amer. Statist. Ass., 74, 537-552, 1979b.

HABERMAN, S.J., Log-linear models for frequency tables with ordered classifications, Biometrics, 30, 589-600, 1974.

McCULLAGH, P., A class of parametric models for the analysis of square contingency tables with ordered categories, Biometrika, 65, 413-418, 1978.

McCULLAGH, P., Regression models for ordinal data (with discussion), J. Roy. Statist. Soc. (B), 42, 109-142, 1980.

NISHISATO, S., Analysis of categorical data: Dual scaling and its applications, University of Toronto Press, Toronto, 1980.

SIMON, G., Alternative analyses for the singly ordered contingency table, J. Amer. Statist. Ass., 69, 971-976, 1974.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report #255	2. GOVT ACCESSION NO. AD-A121923
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Using Information on Ordering for Loglinear Model Analysis of Multidimensional Contingency Tables	5. TYPE OF REPORT & PERIOD COVERED
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Stephen E. Fienberg	8. CONTRACT OR GRANT NUMBER(s): N00014-80-C-0637
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Contracts Office, Carnegie-Mellon University, Pittsburgh, PA 15213	12. REPORT DATE: June, 1982
	13. NUMBER OF PAGES: 9
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Categorical data; Loglinear models; Maximum likelihood estimates; Multidimensional contingency tables; Ordered categories; Ordinal variables.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Many different authors have claimed that the loglinear model approach to the analysis of contingency table data is appropriate only for nominal variables and does not make use of information on the ordinal nature of some categorical variables (i.e. the ordering of the categories). In this paper, we review a variety of loglinear model methods which do take into account, either explicitly or implicitly, such information on ordering. Our focus is on methods involving maximum likelihood estimation, but other methods of estimation can be used with these models. We also consider briefly, some additional models for ordered categorical data.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)