J Natn Sci Foundation Sri Lanka 2007 35(1) 29-40

RESEARCH ARTICLE

Log-linear models for ordinal multidimensional categorical data

Prabhani Kuruppumullage and Roshini Sooriyarachchi*

Department of Statistics University of Colombo Colombo 03

Revised 16 October 2006 Accepted 19 January 2007

Abstract This study emphasised on methods for analysis of categorical data having ordered categories for the multidimensional case and the paper discusses some of the specialized models which efficiently use the information on the ordering, unlike standard methods for nominal categorical data, for multidimensional variables In order to illustrate the methodology, three dimensional data from a shopping survey in Oxford was used

The Standard nominal model fitted, represented the associations between the life cycle level, car availability and the agreement with the statement *I find getting to grocery shops very tiring*, with 16 degrees of freedom The model selected taking the ordinal nature of variables into account also represented the same associations with 27 degrees of freedom, thus with lesser number of parameters The standard log-linear model requires describing interactions using a number of parameters where as when ordinal nature of the variables is considered, interactions can be represented by a few parameters

Based on the model which takes into consideration the ordinal nature of the variables the odds ratios to illustrate the association between the life cycle and agreement, disagreement, tendency to disagree, in-between, and tendency to agree with statement are 0 8, 0 4, 0 9 and 0 9 respectively The odds ratio that describes the association of the car availability and the agreement with the statement is 0 91

It is established that ordering of categories utilizes the information reflected from data where as nominal models do not use the information in the ordered categories Also the suggested models have less parameters and are thus simpler and more parsimonious

Key Words: Column effects model, Linear-by-linear association model, Log-linear model, Ordinal categorical data, Row-effects model

INTRODUCTION

Statistical methodology to analyse categorical data has only recently reached the level of sophistication achieved early in this century by methodology for analysis of continuous data[1] The recent development of methods for analysis of categorical data was stimulated by increasing methodological sophistication in the social and bio-medical sciences[1]

In analyzing categorical data it is necessary to consider the scale of measurement and there are two main types of scales of interest Nominal scale is the simplest scale of measurement which assumes that distinct levels differ in quality but not in quantity and ordinal scale considers the difference of distinct levels and even a hierarchy of importance

Ordinal scales are pervasive in the social sciences, in particular, for measuring attitudes and opinions on various issues and states of various response types Besides they occur commonly in many diverse fields[1] In many studies, ordinal variables are analyzed using nominal techniques assuming that results are invariant to permutations of the categories of any of the variables[2] This sacrifices a certain amount of information when the measurements are of ordinal scale Because of this, for many years in the past, researches have been carried out to extend the log-linear models to perform more complete and informative analysis for ordinal data It has been proposed that there are many advantages to be gained from using ordinal methods instead of, or in addition to the standard nominal procedures[2] Ordinal methods represent two-way associations using a single parameter whereas in standard nominal case, summarization of tables is required As ordinal methods have lesser number of parameters compared to nominal models, they are more parsimonious and thus have more power to test the significance of the interaction terms in the log-linear model (Proof give in Annex 2)

In this paper an attempt has been made to view the importance of utilizing the ordinal property of ordinal

*Corresponding author

variables in the analysis of categorical data and determine if any differences occur in the results obtained from these ordinal models when compared to standard categorical data techniques that treat all the variables as nominal.

Application of these theories have been discussed by several authors[1,2,3,4] only for two-way tables, where as in this paper attention is paid to application of these theories to the multi-dimensional case. In order to demonstrate the usage of orderings of categories, data from a Shopping Survey in Oxford[5] was used. There were three variables of interest namely, the life cycle levels, car availability and the agreement with the statement *'I find getting to grocery shops very tiring'*. Section three provides a description of the levels of each of these three variables. This three dimensional example can easily be generalized to the multidimensional case.

The objectives of this study were to explore the advantages of considering the ordering of levels of categorical variables with respect to:

i. Obtaining simpler models which are easily interpretable
ii. Providing easier quantification of associations
iii. Improving the power to test the significance of the interaction terms in the log-linear model

Section 2 of this paper introduces the theory behind the methods used for the study and section 3 provides an illustration using an example. Section 4 discusses the results and draws conclusions. Annex 1 provides some vital proofs and tables required to understand the usage of considering the ordinal nature of variables and Annex 2 gives a proof that ordinal models have more power to detect interactions when compared to nominal models.

METHODS AND MATERIALS

The main aspect of this study was to explore the advantages of considering the ordering of levels of categorical variables.

In order to visualize the above mentioned advantages it is necessary that the data is analyzed in both ways, by treating levels as nominal and by considering the natural orderings of the levels.

The models that were used in the analysis were as follows.

a. Standard log-linear model[6].
b. Linear-by-Linear Association in Two-Way Tables[2].
c. Row Effects Model[2].

Standard Log-Linear Model[5] **:** The form of the log-linear model under the null hypothesis, H_0: Independence of two variables X and Y for the simplest case of two binary variables in the case of a $I \times J$ contingency table is

$$log_e(e_{ij}) = \mu + \lambda_i^X + \lambda_j^Y; \; i = 1, \ldots, I \; and \; j = 1, \ldots J \; ......(1)$$

where

μ denotes the overall mean, λ_i^X denotes the effect of i^{th} level of X, and λ_j^Y denotes the effect of j^{th} level of Y. e_{ij} is the expected value of the cell formed by the i^{th} level of X and the j^{th} level of Y.

and the saturated model[2] is

$$log_e(e_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}; \; i = 1, \ldots, I \; and \; j = 1, \ldots, J \ldots\ldots(2)$$

Here the terms with two subscripts pertain to partial associations between the corresponding variables. For the standard log-linear model, independence is given by the null hypothesis H_0: $\lambda_{ij}^{XY} = 0$ *for all* $i = 1, \ldots, I$ *and* $j = 1, \ldots, J$.

The likelihood ratio test statistic for testing H_0 is $L_R = G_1^2 - G_2^2$. The notation G_k^2 denotes the deviance of model k where $k=1, 2$. Under H_0, L_R has an asymptotic chi-square distribution with $(I-1)$ $(J-1)$ degrees of freedom. This result can be used to test H_0.

Linear-by-linear association in two-way tables[2]: For two-way tables one rarely expects the independence model to fit well. For a model to have much scope it must allow association, yet retain some residual degrees of freedom, that is, it nests between the independence model and the saturated model. The linear-by-linear model is a simple model of this type for association between two ordinal variables.

The model requires assigning scores $\{u_i\}$ and $\{v_j\}$ to the rows and columns. To reflect category orderings we take $u_1 \leq u_2 \leq \ldots\ldots u_I$ and $v_1 \leq v_2 \leq \ldots\ldots v_J$. It is then possible rather than going directly to the fully saturated model to explore the model which has an interaction structure that directly reflects the ordering of the rows and columns and of the scores $\{u_i\}$ and $\{v_j\}$.

The linear-by-linear association model is then

$$log_e(e_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \beta u_i v_j \; ; \; i = 1, \; .., I \; and \; j = 1, \; .., J \; . ....(3)$$

Where e_{ij} is the expected value for the cell made up by the i^{th} row and j^{th} column, μ represents the overall mean effect, λ_i^X represents the effect of the i^{th} level of the variable

X, λ_j^Y represents the effect of the j^{th} level of the variable Y. Parameter β describes the association between X and Y. Values u_i and v_j are the known scores assigned to the rows and columns. Often the above model is taken as,

$$log_e(e_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \beta(u_i - \hat{u})(v_j - \overline{v}) \qquad ......(4)$$

which is known as the *centered model*.

The independence model is a special case when $\beta = 0$. Since $\{u_i\}$and $\{v_j\}$are fixed the linear-by-linear association model has only one more parameter (β) than the independence model. Thus the degrees of freedom for the linear-by-linear association model are

$$df = IJ - [1 + (I-1) + (J-1) + 1] = IJ - I - J$$

Note that model (3) is unsaturated for tables with $J>2$. It is a special case of the saturated model (2), in which λ_{ij}^{XY} takes the form $\beta u_i v_j$. While the linear-by-linear association model requires one parameter (β) to describe association regardless of I and J the saturated model requires $(I-1)(J-1)$ parameters. In many applications the choice of scores will reflect assumed distances between midpoints of categories for an underlying interval scale. Equally spaced scores result in the simplest interpretation for the model discussed in this section. In practice, the integer scores $\{u_i = i\}$and $\{v_j = j\}$are most commonly used.

Row effects model^2 *:* Here the row variable *(X)* is treated as nominal and column variable *(Y)* as ordinal. The model is appropriate for two-way tables with ordered column classifications. The ordered scores $v_1 \leq v_2 \leq ......v_J$ are assigned to reflect the ordering of the columns. The rows are now unordered (nominal). Using the linear-by-linear structure as before and replacing the ordered values $\{\mu_i\}$ in the linear-by-linear association model by the unordered parameters gives the row effects model

$$log_e(e_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \mu_i v_j \qquad ......(5)$$

The degrees of freedom for this model is *df = total cells – number of independent parameters = (I-1)(J-2)*

Direct ordinal test of independence: For the linear-by-linear association model, independence is given by the null hypothesis H_0: $\beta = 0$. The likelihood ratio test statistic for testing H_0 is, $G^2(I|L \times L) = G^2(I) - G^2(L \times L)$. The notation $G^2(I)$ represents the deviance of the standard nominal model and $G^2(L \times L)$ represents the deviance of the linear-by-linear model. The likelihood ratio statistics $G^2(I|L \times L)$ is the difference of these two deviances and measures to what extent the nominal model is a better fit compared to the linear-by-linear model. When the $L \times L$ model holds, the ordinal test using $G^2(I|L \times L)$ is asymptotically more powerful than the test using $G^2(I)$ (Proof give in Annex 2). The power of a chi-squared test increases when degree of freedom decreases, for fixed non-centrality. When the $L \times L$ model holds, the non-centrality is the same for $[G^2(I|L \times L)]$ and $G^2(I)$. Thus, $G^2(I|L \times L)$ is more powerful.

Likelihood Equations and Model Fitting: The Poisson log-likelihood $L(\underset{\sim}{\theta}) = \Sigma_i\Sigma_j n_{ij} \, log_e e_{ij} - \Sigma_i\Sigma_j e_{ij}$ simplifies for the $L \times L$ model $log_e(e_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \beta u_i v_j$ to

$$L(\underset{\sim}{\theta}) = n\mu + \Sigma_i n_{i+} \lambda_i^X + \Sigma_j n_{+j} \lambda_j^Y + \beta\Sigma_i\Sigma_j u_i v_j n_{ij} -$$
$$\Sigma_i\Sigma_j \, exp\,(\mu + \lambda_i^X + \lambda_j^Y + \beta u_i v_j) \qquad ...... (6)$$

Here$\underset{\sim}{\theta}$ is the parameter vector $(\lambda_1^X......\lambda_I^X, \lambda_1^Y......\lambda_J^Y, \beta)$. Differentiating $L(\underset{\sim}{\theta})$ with respect to $(\lambda_i^X, \lambda_j^Y, \beta)$ for $i = 1,... I$, $j = 1,... J$ and setting the three partial derivatives equal to zero yields likelihood equations

$$\hat{e}_{i+} = n_{i+} \qquad ; i = 1, ..., I \qquad ...(7)$$

$$\hat{e}_{+j} = n_{+j} \qquad ; j = 1, ..., J \qquad ...(8)$$

$$\Sigma_i\Sigma_j u_i v_j \hat{e}_{ij} = \Sigma_i\Sigma_j u_i v_j n_{ij} \qquad ...(9)$$

Where a (+) sign corresponds to summing over the corresponding suffix and a hat sign (^) corresponds to the maximum likelihood (ML) estimate.

Iterative methods such as Newton-Raphson yield the ML fit.

Let $p_{ij} = \dfrac{n_{ij}}{n}$ *and* $\hat{\pi}_{ij} = \dfrac{\hat{e}_{ij}}{n}$

The third likelihood equation *{equation (9)}* implies that

$$\Sigma_i\Sigma_j u_i v_j \hat{\pi}_{ij} = \Sigma_i\Sigma_j u_i v_j p_{ij} \qquad ...(10)$$

Since marginal distributions and marginal means and variances are identical for fitted and observed distributions, the equation (9) implies that the correlation between the scores for X and Y is the same for both distributions. The fitted counts display the same positive or negative trend as the data.

Since $\{u_i\}$ and $\{v_j\}$ are fixed, the $L \times L$ model $log_e(e_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \beta u_i v_j$ has only one more parameter *(β)* than the independence model. Its residual

$df = IJ - I - J$ is unsaturated for all other tables except for 2×2 tables.

Ordinal variables in models for multi-way tables: Generalizations of association models can be used for multidimensional tables with ordinal responses. In three dimensions, the rich collection of models includes

1. association models that are more parsimonious than the nominal model *(XY, XZ, YZ).*
2. models permitting heterogeneous association that, unlike model *(XYZ)*, are unsaturated.

Models for association that are special cases of *(XY, XZ, YZ)* replace λ association terms by structured terms that account for ordinality. For instance, when both *X* and *Y* are ordinal, alternatives to λ_{ij}^{XY} are a linear-by-linear term $\beta u_i v_j$, a row effects term $\mu_i v_j$, or a column effects term $u_i \lambda_j$; these provide a stochastic ordering of conditional distributions within rows and columns, or just within rows, or just within columns. With a linear-by-linear term the model is,

$$\log_e e_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \quad ...(11)$$

$$i = 1, ..., I; \qquad j = 1, ..., J; \qquad k = 1, ..., K$$

The conditional odds ratio $\theta_{ij(k)}$ then satisfies

$$\log_e \theta_{ij(k)} = \beta\, (u_{i+1} - u_i)\, (v_{j+1} - v_j) \textit{ for all } k$$

The u_i's and v_j's are as defined before.

The association is the same in different partial tables, with homogeneous linear-by-linear XY association. When the association is heterogeneous, structured terms for ordinal variables make effects simpler to interpret than in the saturated model. For instance, the heterogeneous linear-by-linear association XY model

$$\log_e e_{ijk} = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta_k u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \quad ...(12)$$

allows the *XY* association to change across levels of *Z*. With unit-spaced scores,

$$\log_e \theta_{ij(k)} = \beta_k \textit{ for all } i \textit{ and } j$$

It has uniform association within each level of Z, but heterogeneity among levels of Z in the strength of association. Fitting it corresponds to fitting the $L \times L$ model separately at each level of *Z*.

3. An example

In order to illustrate the methodology, data from a Shopping Survey in Oxford[4] was used. Three ordinal variables were considered for the illustration and those were life cycle levels, car availability, and agreement with the statement *'I find getting to grocery shops very* tiring'. The variable *life cycle levels* has three levels as middle-aged, younger people with children and younger people without children. The variable *car availability* also has three levels as no car availability, some car availability and full car availability. The third variable has five levels as disagree, tend to disagree, in between, tend to agree, and agree.

In order to visualize the advantages of considering the natural ordering of variables, first the standard log-linear model was obtained using the forward selection technique[2]. SAS *PROC CATMOD* was used to obtain the standard log-linear model and *PROC* GENMOD was used when ordinal nature is considered. The respective constraints for these procedures were sum of parameters equals zero and parameter for last level equals zero.

The selected standard nominal model to represent the associations was as follows.

Model A

$$\log_e (e_{ijk}) = const + (life)_i + (car)_j + (agree)_k + (car{*}agree)_{jk} + (life{*}agree)_{ik} + (life{*}car)_{ij} \quad ...\ (13)$$

Using the likelihood ratios, goodness of fit of the above model was assessed. The corresponding hypothesis is;

H_0: *Selected model represents associations well enough*

vs

H_1: *Selected model does not represent associations well enough*

The deviance of the chosen model is 24.2046. Under H_0 the deviance has an asymptotic chi-square distribution with 16 degrees of freedom. Since the calculated deviance is less than $\chi^2_{16,\ 5\%}$ (= 26.2962), the null hypothesis of well fitted model should not be rejected at 5% significance level. The p-value of the selected model when compared with the saturated model is 0.0851 which implied that the chosen model represent the associations well. Hence the best standard log-linear model to represent the associations between the variables *life cycle levels, car availability and agreement with the statement* was taken to be model A.

It is clear from the selected standard log-linear model that all variables *(car availability, life cycle levels, and agreement with the statement)* are associated. However these associations do not change with the level of the third variable.

After obtaining a model to represent the associations between the variables by ignoring the natural orderings of the variables it is of interest to fit a model which utilizes the natural orderings of the variables.

As the *backward elimination procedure*[2] is to be used in selecting the most representative model by utilizing the natural orderings of the variables, the model fitting was started by taking the selected standard model into account. The strategy adopted in model selection was to first examine whether any nominal terms can be replaced by linear by linear interaction terms and second to examine whether remaining nominal terms can be replaced by row-effect or column-effect terms. In doing this two things are considered.

Initially the term resulting in the largest p-value *(>0.05)* for the difference in deviance is selected and then the goodness of fit of this selected model is examined (using p-value). Only if both the p-values of the difference in deviance and goodness of fit are not significant *(>0.05)* is the relevant term selected.

*Selection of linear-by-linear interaction term(s):*When considering model A it is clear that the first two-factor interaction term that was chosen to be treated as linear was *car*agree*. It was found that both these ordinal variables could be treated as linear without making any significant changes to the goodness of fit of the selected standard log-linear model. The results obtained in each of the cases are tabulated in table 3.1. The notation $s_i^{life} = i$ represents the scores associated with the variable *life(life cycle)* when treated as linear, $u_j^{car} = j$ represents the scores associated with the variable *car(car availability)* when treated as linear and $v_k^{agree} = k$ represents variable *agree(agreement with the statement)* when treated as linear.

When the deviance increments with respect to model A were considered it was seen that, the model in which both the variables *car* and *agree* were treated as linear resulted in the largest p-value *(0.1502)* and is greater than 0.05. Thus it is possible to treat both variables in the term *car*agree* in model (12) as linear. The p-value of the model for the goodness of fit is 0.053 which is marginally higher than 0.05. This shows that while the fit of the model is not excellent it is adequate.

Model B:

$$\log_e(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k + [(u_j^{car} * v_k^{agree})] + (life*agree)_{ik} + (life*car)_{ij} \qquad \ldots(14)$$

Then the next two-way interaction terms were treated as linear-by-linear terms and the results are summarized in table 3.2.

It is revealed from the Table 3.2 that the largest p-value corresponding to the difference is 0.0894 in model in which both variables *life(life cycle)* and *agree (agreement with the statement)* are treated as linear. This

Table 3.1: Summary statistics obtained by treating each of the interactions linear in model A

Model	Deviance	Degree of freedom	Difference with model A		p-Value of difference
			Deviance	DF	
Model A	24.2046	16	-	-	-
$\log_e(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k$ $+ (u_j^{car} * v_k^{agree}) + (life*agree)_{ik} + (life*car)_{ij}$	34.9491	23	10.7445*	7	0.1502
$\log_e(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k$ $+ (car_j * agree_k) + (s_i^{life} * v_k^{agree}) + (life*car)_{ij}$	35.8763	23	11.6717*	7	0.1119
$\log_e(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k$ $+ (car_j * agree_k) + (life*agree)_{ik} + (s_i^{life} * u_j^{car})$	36.4599	19	12.2553	3	0.0066

*- *Deviance increment not significant at 5% level*

Table 3.2: Summary statistics obtained by treating each of the interactions linear in model B

Model	Deviance	Degree of	Difference with model B		p-Value of difference
	freedom		Deviance	DF	
Model B	34.9491	23	-	-	-
$\log_e(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k$ $+ (u_j^{car} * v_k^{agree}) + (s_i^{life} * v_k^{agree}) + (life*car)_{ij}$	47.3077	30	12.3586*	7	0.0894
$\log_e(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k$ $+ (u_j^{car} * v_k^{agree}) + (life*agree)_{ik} + (s_i^{life} * u_j^{car})$	47.5162	26	12.5671	3	0.0057

*- *Deviance increment not significant at 5% level*

Table 3.3: Summary statistics obtained by treating each of the interactions row-effects/column-effects in model B

Model	Deviance	Degree of freedom	Difference with model A		p-Value of difference
			Deviance	DF	
Model B	34.9491	23	-	-	-
$\log_e(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k$ $+ (u_j^{car} * v_k^{agree}) + (s_i^{life} *agree_k) + (life*car)_{ij}$	39.2052	27	4.2561*	4	0.3725
$\log_e(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k$ $+ (u_j^{car} * v_k^{agree}) + (life_i * v_k^{agree}) + (life*car)_{ij}$	43.5848	29	8.6357*	6	0.1951
$\log_e(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k$ $+ (u_j^{car} * v_k^{agree}) + (life*agree)_{ik} + (s_i^{life} *car_j)$	44.3399	25	9.3908	2	0.0091
$\log_e(e_{ijk}) = const + (life)_i + (car)_j + (agree)_{ks}$ $+ (u_j^{car} * v_k^{agree}) + (life*agree)_{ik} + (life_i * u_j^{car})$	39.3906	25	4.4415*	2	0.1085

*- *Deviance increment not significant at 5% level*

value is greater than 0.05 and thus the model is not significantly different from model B. But it is to be noticed that the goodness of fit p-value of this model is 0.0232 which implies that the model does not fit well. Thus this model was not chosen.

Thus it was concluded that the only linear-by-linear term that could be included in the model is $(u_j^{car} * v_k^{agree})$.

Selection of row-effects/column effects interaction term(s): After selecting the linear-by-linear terms, it was attempted to make other two-way interaction terms row-effects/column-effects by using the same strategy used in the above section.

The two-way interaction terms *life*agree* and *life*car* were treated as row-effects and column-effects terms respectively and the results are summarized Table 3.3.)Table 3.3 reveals that the largest p-value corresponding to deviance difference is 0.3725 which is greater than 0.05. This is when variable *life(life cycle)* is treated as linear and variable *agree(agreement with the statement)* is treated as factor in the two-factor interaction *life*agree*. Thus it could be concluded that two-way interaction *life*agree* could be treated as row-effects term. When the goodness of fit p-value of the model *(=0.0607)* is considered it is clear that this model fits well at 5% significance level. Thus the selected model is as follows.

Model C:

$$log_e\,(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k + (u_j^{car} * v_k^{agree}) + (s_i^{life} * agree_k) + (life * car)_{ij} \qquad ...(15)$$

The similar procedure was applied to obtain more row-effects/column-effects terms and the results are summarized in Table 3.4.

The largest p-value corresponding to deviance difference is 0.1088 which is greater than 0.05. Thus the model is not significantly different from model C. But when the goodness of fit of the model is considered it is seen that this model does not fit well as the p-value corresponding to the model *0.0397* is less than *0.05*. Thus this model is not selected and it is concluded that model C is the best model obtained by taking the natural orderings of the variables into account.

When referring the model C it could be seen that the model C is a combination of a linear-by-linear association term $(u_j^{car} * v_k^{agree})$, a row-effects term $(s_i^{life} * agree_k)$ and two-way nominal interaction term between $(life_i * car_j)$.

After selecting the model, it is necessary to assess the goodness of fit of the selected model. The following hypothesis is used for this.

H_0: *Selected model represents associations well enough*

vs

H_1: *Selected model does not represent associations well enough*

The deviance of the selected model is 39.2052 with 27 degrees of freedom. This value is compared with the corresponding chi-square table value $\chi^2_{27,\,5\%}(= 40.1133)$ And it is seen that the value of the deviance of the chosen model is less than the corresponding chi-square table value. Also the p-value of the selected model is *0.0607* (> *0.05*) which implied that the chosen model represents associations well. Thus it is possible to conclude at the 5% significance level that there is no sufficient evidence to say that the chosen model does not fit well enough. Hence the best model to represent the associations between factors life cycle level, car availability, and agreement with the statement *'I find getting to grocery shops very* tiring' after considering the natural orderings of the appropriate variables is model C.

Also it is necessary that the selected model does not deviate significantly from the standard log-linear model. Thus a comparison of the standard log-linear model and the combination model obtained by considering the natural ordering were compared.

Table: 3.4 - Summary statistics obtained by treating each of the interactions row-effects/column-effects in model C

Model	Deviance	Degree of freedom	Difference with model B		p-Value of difference
			Deviance	DF	
Model C	39.2052	27	-	-	-
$log_e\,(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k + (u_j^{car} * v_k^{agree}) + (s_i^{life} * agree_k) + (s_i^{life} * car_j)$	47.3058	29	8.1006	2	0.0175
$log_e\,(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k + (u_j^{car} * v_k^{agree}) + (s_i^{life} * agree_k) + (life_i * u_j^{car})$	43.6409	29	4.4357*	2	0.1088

*- *Deviance increment not significant at 5% level*

The standard log-linear model obtained in this paper is as follows.

Model A:

$$\log_e(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k + (car*agree)_{jk} + (life*agree)_{ik} + (life*car)_{ij}$$

The combination model obtained in this paper by considering the ordinal property of the appropriate variables is as follows.

Model C:

$$\log_e(e_{ijk}) = const + (life)_i + (car)_j + (agree)_k + (u_j^{car} * v_k^{agree}) + (s_i^{life} * agree_k) + (life*car)_{ij}$$

To assess whether the consideration of natural orderings of variables increase the deviance of the standard log-linear model, comparison of the model A and model C was done and the results are tabulated in table 3.5.

It is also known that this deviance increment follows an asymptotic chi-square distribution and thus the corresponding table value $\chi^2_{11,5\%}(= 19.6751)$ is used to assess the following hypothesis.

H_0: There is no difference between the two models

vs

H_1: There is a difference between the two models

As the deviance increment due to consideration of linearity of possible variables *(=15.0006)* is less than the corresponding chi-square table value $\chi^2_{11,5\%}(= 19.6751)$, it is concluded at 5% significance level not to reject the null hypothesis. Thus it could be concluded at 5% significance level that there is no sufficient evidence to say that the two models are significantly different with respect to the fit. Hence it is possible to select the *model C* which had higher number of degrees of freedom and thus a simpler model. The whole idea behind this paper is to utilize information revealed by the natural orderings of the variables. And it was discovered that by considering the natural ordering of the variables, it is possible to save 11 degrees of freedom and could thus obtain a simpler model compared to the standard log-linear model.

Parameter estimation, odds ratio calculation and model interpretation: After selecting the model the interest is then to interpret the model using the parameter estimates and appropriate odds ratios. For two-way interactions where both the variables are treated as factors, it is attempted to look at the estimated parameters for a better interpretation.

Table 3.5: Summary statistics obtained in comparing *model A* with *model C*

Model	Deviance	Degree of freedom	Difference with model 1		p-Value of difference
			Deviance	DF	
Model A	24.2046	16	-	-	-
Model C	39.2052	27	15.0006*	11	0.1825

*- *Deviance increment not significant at 5% level*

Table 3.6: Parameter estimates of two-factor association *life cycle levels & car availability* .

		Life cycle levels		
		Middle aged	(younger people) with children	(younger people) without children
Car Availability	No car availability	0.1475	-0.4621	0.0000
	Some car availability	0.5022	0.1509	0.0000
	Full car availability	0.0000	0.0000	0.0000

It is seen that middle-aged people had a high chance to have some car availability compared to younger people without children. Also the chance of having no car availability for younger people with children is lower than the younger people without children.

To interpret the associations with linear effects relative odds ratios are calculated and have been explained below. As in the two-way interaction *life*agree,* it is found that variable *life* behaves linearly and the corresponding odds ratio is obtained for this case.

The estimates corresponding to first four levels of the agreement with the statement with respect to the respondents who had agreed with the statement are $\hat{agree}_1 = 0.2316$, $\hat{agree}_2 = 0.8061$, $\hat{agree}_3 = 0.0092$ *and* $\hat{agree}_4 = 0.1273$. The further $\hat{agree}_i$ falls in the positive direction, the greater the tendency for the respondents with level of agreement *i* to locate at the maximum life cycle level (i.e: younger people without children) relative to respondents who have agreed to the statement. In this case younger people without children have disagree with the statement *'I find getting to grocery shops very tiring'* than the younger people with children and middle-aged people.

The derivation of the estimated odds ratios are given in the Annex 1. All estimated log odds ratios are negative indicating a tendency for middle aged people to agree with the statement.

When $k=1$;

The estimated odds that a respondent who has agreed with the statement being younger person without children instead of younger person with children, or being younger person with children instead of middle-aged, are 0.8 times $\{exp(-0.2316) = 0.7933\}$ the corresponding estimated odds for a respondent who has disagreed with the statement. The 95% confidence interval is *(0.6274, 1.0030)*. As one is included in the confidence interval, null hypothesis $\theta = 1$ is not rejected. This indicates that there is no difference in odds, of being younger person without children instead of younger person with children, or being younger person with children instead of middle-aged, between respondents who have agreed and disagreed with the statement.

However here the upper limit is only just above one and therefore the result is nearly significant Thus, it could be concluded that the estimated odds that a respondent who finds getting to grocery shops very tiring being younger person without children instead of younger person with children, or being younger person with children instead of middle-aged is around 0.8 times less compared to a respondent who does not find getting to grocery shops very tiring.

When $k=2$;

The estimated odds that a respondent who has agreed with the statement being younger person without children instead of younger person with children, or being younger person with children instead of middle-aged, are 0.4 times $\{exp(-0.8061) = 0.4466\}$ the corresponding estimated odds for a respondent who has tend to disagreed with the statement. The 95% confidence interval is *(0.2773, 0.7194)*. As one is not included in the confidence interval, null hypothesis $\theta = 1$ is rejected. Thus the confidence interval supports the conclusions taken using odds ratio. Thus it is concluded that the estimated odds that a respondent who finds getting to grocery shops very tiring being younger person without children instead of younger person with children, or being younger person with children instead of middle-aged is less (0.4 times) comparatively to a respondent who does not tend to find getting to grocery shops very tiring.

When $k=3$;

The estimated odds that a respondent who has agreed with the statement being younger person without children instead of younger person with children, or being younger person with children instead of middle-aged, are 0.9 times $\{exp(-0.0092) = 0.9908\}$ the corresponding estimated odds for a respondent who has responds (agreed) in-between with the statement. The 95% confidence interval is *(0.6756, 1.4532)*. As one is included in the confidence interval, null hypothesis $\theta = 1$ is not rejected. This indicates that there is no difference in odds of being a younger person without children instead of younger person with children or being younger person with children instead of middle-aged between respondents who have agreed and are in-between.

When $k=4$;

The estimated odds that a respondent who has agreed with the statement being younger person without children instead of younger person with children, or being younger person with children instead of middle-aged, are 0.9 times $\{exp(-0.1273) = 0.8805\}$ the corresponding estimated odds for a respondent who has tend to agree with the statement. The 95% confidence interval is *(0.5978, 1.2967)*. As one is included in the confidence interval, null hypothesis $\theta = 1$ is not rejected. The interpretation is similar to $k=3$.

Also it is found that both the variables in the two-way interaction term *car*agree* where *car* represents the car availability and *agree* represents the agreement with the statement, could be treated as linear. Thus the selected term is a linear-by-linear term and the information reflected from this term could be interpreted as follows.

The *ML* estimate $\hat{\beta} = -0.0950$ (negative) indicates that the respondents having higher car availability tend to disagree more with the statement *'I find getting to grocery shops very tiring'*.

The corresponding odds ratio is;

$$\hat{\theta}_{ik(j)} = exp\ (\hat{\beta}) = exp\ (-0.0950) = 0.9094$$

Thus the estimated odds that a respondent agreeing with the statement (who finds getting to grocery shops very tiring) having full car availability instead having some car availability, or having some car availability instead of no car availability is approximately 0.9 times $\{exp\ (-0.0950) = 0.9094\}$ the corresponding estimated odds for a respondent who does not find getting to grocery shops very tiring. The 95% confidence interval is *(0.8615, 0.9599)* which does not include 1. Thus the null hypothesis $\theta = 1$ is rejected and hence it is seen that the confidence interval supports the conclusions taken using the odds ratio.

Thus it could be concluded that respondents do not find getting to grocery shops tiring with increasing car availability.

DISCUSSION

The main objective of this paper was to discuss the advantages of using the natural orderings of ordinal categorical variables. It was of particular interest to illustrate that models which use the natural orderings are simpler, provide easier quantification of associations in terms of odds ratios and have more power to detect interactions when compared to models which use nominal scale variables. A three-dimensional example was used for illustration.

A standard log-linear model was chosen to explore the associations among life cycle levels, car availability, and agreement with the statement *'I find getting to grocery shops very tiring'* by considering the variables to be of nominal scale.

Then the model was chosen by taking the natural orderings of the variable into account. In this case it was found that in the association between *car* and *agree*, both variables *car* and *agree* could be treated as linear variables than factors. And also in the association between *life* and *agree*, variable *life* could be treated as continuous. This reduced the number of parameters corresponding to each of the interactions. Due to this parameter reduction the degrees of freedom corresponding to the model was increased (by 11) and thus it was possible to obtain a more parsimonious model with much simpler interpretations.

Estimated odds showed that, respondents with higher car availability do not find getting to grocery shops very tiring and thus tends to disagree with the statement *'I find getting to grocery shops very* tiring'.

Also respondents who have agreed with the statement being younger persons without children instead of younger persons with children, or being younger persons with children instead of being middle-aged, are 0.4 times $\{exp(-0.8061) = 0.4466\}$ the corresponding estimated odds for a respondent who tends to disagree with the statement.

It was clear from the analysis that the suggested model tests the associations with 27 degrees of freedom where as standard log-linear model tests the same associations only with 16 degrees of freedom. Thus the number of parameters used to interpret the associations in the suggested model is less than in the standard log-linear model, illustrating that the suggested model is simpler. Though standard log-linear models require 2×2 sub-tables to describe the interactions, the suggested model can be easily utilized to calculate odds ratios (Annex 1) to describe the similar interactions. As shown in Annex 2 the power for testing associations is higher in the suggested model.

Throughout this paper an illustration of methods of selection of terms, deciding the form of the terms and interpreting terms have been studied for the three dimensional case. This could be easily generalized, using the same approach for the multi dimensional case. This work could further be extended by examining the magnitude of increase in power of the likelihood-ratio tests, when the ordinal nature of the categorical variables is utilized, by using simulation studies.

References .

1. Agresti A. (1984). *Analysis of Ordinal Categorical Data*, John Wiley and Sons, New York.
2. Agresti A. (2002). *Categorical Data Analysis*, John Wiley and Sons, New York.

3. Fingleton B. (1984). *Models for Category Counts*, Cambridge University Press. UK.
4. Clogg C. & Shihadeh E. (1994). Statistical Models for Ordinal Variables, Thousands Oaks, CA, Sage Publications. Review Author: Agresti A. (1995). *Contemporary Sociology* **24**(5):711-712.
5. Bowly S. & Silk J. (1982). Analysis of qualitative data using GLM: two examples based on shopping survey data. *The Professional Geographer* **34**: 80-90.
6. Dobson A.J. (2002). *An Introduction to Generalized Linear Models*, Second Edition, Chapman and Hall.
7. Patnaik P.B. (1949). The Non-Central x^2 and F distributions and their applications. *Biometrika* **36**: 202-232.

Annex 1

The derivation of the estimated odds ratios;

$$log_e(\hat{\theta}_{ik(j)}) = log_e\left[\frac{[(\mu_{(i+1)j(k+1)})*(\mu_{ijk})]}{(\mu_{(i+1)jk})*(\mu_{ij(k+1)})}\right]$$

$$= log_e(\mu_{(i+1)j(k+1)}) + log_e(\mu_{ijk}) - log_e\mu_{(i+1)jk} - log_e\mu_{ij(k+1)}$$

By substituting the values;

$$log_e(\hat{\theta}_{ik(j)}) = \left\{\begin{array}{l} [const + (life)_{(i+1)} + (car)_j + (agree)_{(k+1)} + \\ (life*car)_{(i+1)j} + ((i+1)*agree)_{(k+1)} + ((i+1)*j)] + \\ \\ [const + (life)_i + (car)_j + (agree)_k + \\ (life*car)_{ij} + (i*agree)_k + (i*j)] - \\ \\ [const + (life)_{(i+1)} + (car)_j + (agree)_k + \\ (life*car)_{(i+1)j} + ((i+1)*agree)_k + ((i+1)*j)] - \\ \\ [const + (life)_i + (car)_j + (agree)_{(k+1)} + \\ (life*car)_{ij} + (i*agree)_{(k+1)} + (i*j)] \end{array}\right\}$$

Thus;

$$log_e(\hat{\theta}_{ik(j)}) = \{((i+1)*agree)_{(k+1)} + (i*agree)_k - ((i+1)*agree)_k - (i*agree)_{(k+1)}\}$$

$$= (agree)_{k+1} - (agree)_k$$

100(1 - α)% confidence interval for the above is;

$$= \exp[log_e(\hat{\theta}_{ik(j)}) \pm \{Z_{\alpha/2} * \sqrt{(\mathrm{var}(log_e(\hat{\theta}_{ik(j)})))}\}]$$

Thus when *k=1*;

$$log_e(\hat{\theta}_{ik(j)}) = \{(agree)_{(k+1)} - (agree)_k\}$$

$$= [0.0000 - (0.2316)]$$

$$= -\ 0.2316$$

To assess the 95% confidence interval;

$$var\ \{log_e(\hat{\theta}_{ik(j)})\} = var(agree_{(k+1)}) + var(agree_k) - [2*cov\ (agree_{k+1}, agree_k)]$$

$$= 0 + 0.0143 - (2*0) = 0.0143$$

Thus 95% confidence interval is;

$$= exp\,[log_e(\hat{\theta}_{ik(j)}) \pm \{Z_{\alpha/2} * \sqrt{\mathrm{var}}\,(log_e(\hat{\theta}_{ik(j)}))\}]$$

$$= exp\ [-0.2316 \pm (1.96 * \sqrt{(0.0143)})]$$

$$= exp\ [-0.2316 \pm 0.2346]$$

$$= exp\ [-0.4662,\ 0.0030]$$

$$= (\mathbf{0.6274,\ 1.0030})$$

When *k=2*;

$$log_e(\hat{\theta}_{ik(j)}) = \{(agree)_{(k+1)} - (agree)_k\} = [0.0000 - (0.8061)]$$

$$= -\ \mathbf{0.8061}$$

Thus 95% confidence interval is;

$$= exp\ [log_e(\hat{\theta}_{ik(j)}) \pm \{Z_{\alpha/2} * \sqrt{(\mathrm{var}\ (log_e(\hat{\theta}_{ik(j)})))}\}]$$

$$= exp\,[-0.8061 \pm (1.96 * 0.2432)] = exp\,[-0.8061 \pm 0.4767]$$

$$= exp\ [-1.2828,\ -0.3294] = (\mathbf{0.2773,\ 0.7194})$$

When *k=3*;

$$log_e(\hat{\theta}_{ik(j)}) = \{(agree)_{(k+1)} - (agree)_k\}$$

$$= [0.0000 - (0.0092)] = -\ \mathbf{0.0092}$$

Thus 95% confidence interval is;

$$= exp\ [log_e(\hat{\theta}_{ik(j)}) \pm \{Z_{\alpha/2} * \sqrt{(\mathrm{var}(log_e(\hat{\theta}_{ik(j)})))}\}]$$

$$= exp\ [-0.0092 \pm (1.96 * 0.1954)]$$

$$= exp\ [-0.0092 \pm 0.3830] = (\mathbf{0.6756,\ 1.4532})$$

When $k=4$;

$$log_e(\hat{\theta}_{ik(1)}) = \{(agree)_{(k+1)} - (agree)_k\}$$
$$= \{0.0000 - (0.1273)\} = \mathbf{-\ 0.1273}$$

Thus 95% confidence interval is;

$$= exp\ [log_e(\hat{\theta}_{ik(j)}) \pm \{Z_{\alpha/2} * \sqrt{}(\mathrm{var}(log_e(\hat{\theta}_{ik(j)})))\}]$$
$$= exp\ [-0.1273 \pm (1.96 * 0.1975)]$$
$$= exp\ [-0.0092 \pm 0.3871] = (\mathbf{0.6728,\ 1.4592})$$

Annex 2

The power of a test is defined to be the probability of rejecting the null hypothesis given that the alternative hypothesis is true.

It is denoted by

$$1 - \beta = Pr\ (rejecting\ H_0\ /\ H_1\ is\ true)$$

For simplicity consider the two dimensional case where variables X and Y are ordinal categorical variables with levels I and J respectively. Consider the three models

$$log\,(e_{ij}) = \mu + \lambda_i^X + \lambda_j^Y \qquad \textit{(model of independence)... (i)}$$

$$log\,(e_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \qquad \textit{(fully saturated model)... (ii)}$$

$$log\,(e_{ij}) = \mu + \lambda_i^X + \lambda_j^Y + \beta u_i v_j \quad \textit{(linear by linear association model)..(iii)}$$

The notation in the models is as defined before. Suppose the three models have deviances D_1, $D_2 = 0$, *and* D_3 respectively. The degrees of freedom of the models are $v_1 = IJ - I - J + 1$, $v_2 = 0$, $v_3 = IJ - I - J$ respectively. The power with which to detect the association between X and Y when such an association is present:

Based on the nominal log-linear model (ii) is given by

$$Power_1 = Pr(D_1 - D_2 > \chi^2_{v_1}\ /\lambda_{ij}^{XY} \neq 0) \qquad ...\ (iv)$$

Based on the log-linear model taking the ordering in to account (iii) is given by

$$Power_2 = Pr(D_1 - D_3 > \chi^2_{v_1 - v_3 = 1}\ /\beta \neq 0) \qquad ...\ (v)$$

Now $DD_1 = D_1 - D_2$ and $DD_2 = D_1 - D_3$ have non-central chi-square distributions with degrees of freedom v_1 and 1 respectively and non-centrality parameter δ_1 and δ_2 respectively[7].

When the linear-by-linear model *{model (iii)}* holds $\delta_1 = \delta_2 = \delta$[1,2]

Then $DD_1 \sim \chi^2_{(v_1,\ \delta)}$

$DD_2 \sim \chi^2_{(1,\ \delta)}$

The power of a chi-squared test increases when degrees of freedom decreases for fixed non-centrality[1,2]. As $v_1 > 1$ for all other cases except for the 2×2 table, Power$_2$ *{given in equation (v)}* is larger than power$_1$ *{given in equation (iv)}* for all other cases except $I = J = 2$. Thus when the linear-by-linear model holds the ordinal test using $G^2(I\,/\,L \times L)$ *{given in equation (v)}* is more powerful than the test using $G^2(I)$ *{given in equation (iv)}*.