Providing an overview of qualitative data sets: Log-linear modelling

Ted Redden

University of New England

Abstract

Papers presented at MERGA conferences in recent years have provided little stimulus for debate on methodological issues relating to research in mathematics education. It seems appropriate that not only should results of research be presented but opportunity for discussion of the processes of research and related issues should be promoted. At the 1993 Brisbane conference of MERGA two papers were presented (Mousley, Sullivan, & Waywood, 1993, Forgasz, Landvogt, & Leder, 1993) that had research methodology as their focus. This focus lay within the qualitative paradigm. This paper continues this theme by considering a strategy for building a conceptual framework for a set of qualitative data and then outlining a procedure for "looking inside" the associations between variables.

Background

This paper describes the use of log-linear analysis in gaining a clearer understanding of a set of categorical data. The analysis of qualitative data is often seen as problematic in that there are few systematic analysis techniques that are available to the researcher. The commonly used statistical procedures such as t-tests, anova, multiple regression etc, require that the data be at the interval or ratio level of measurement. Most research methods texts restrict their discussion of nominal or ordinal data processing procedures to a test of independence of variables using the Pearson Chi squared distribution. While the chi-square test can test a hypothesis of independence it tells us little or nothing about the strength of association. The available measures of association such as contingency coefficient, phi coefficient and Cramer's phi are seen as difficult to interpret and therefore of restricted use. Further the chi square test is of limited use in situations involving a number of variables since a systematic method of investigating the number of interaction effects is elusive.

A systematic method of investigating a set of polychotomous variables where measurement is at the nominal or ordinal level is provided by log-linear modelling. Log-linear modelling became a popular technique in Britain during the 1960's as a technique for analysing medical data. A comprehensive description is provided by Bishop, Fienberg and Holland (1975) while simpler accounts are presented by Everitt (1977) and Tabachnick and Fidell (1989). The calculations involved in the iterative procedures are arduous and hence a computer application is required for effective use of log-linear procedures. Two such applications (in a Macintosh environment) are known to the author. The most complete and difficult to use is the SPSS (Norusis, 1990) platform. A simple to use program is Systat (Wilkinson, 1989), however, this program does not provide parameter estimates and the associated standard errors and, as a result, is of restricted use.

In general two kinds of log-linear analysis can be applied to contingency table data (Haberman, 1978); hierarchical and non-hierarchical. In the hierarchical form, as in multiple regression analysis, if a variable is in the model as a kth order effect it must be in the model as a

(k-1) order effect. In the non-hierarchical form only those main and interaction effects of interest to the hypothesis, and for which a well fitting model can be found, are taken into the analysis. Usually, the hierarchical form is used if the research questions are concerned primarily with measuring the strength of association between categories of the variables, and the non-hierarchical form is employed when the focus of interest is on developing a predictive model. In both cases a search is made for the most parsimonious model that produces a set of expected frequencies which closely match the observed counts (Everitt, 1977). A parsimonious model is one that employs the fewest main and interaction effects. In the present study the principal concern was with measuring strength of association between pairs of components and therefore, an hierarchical approach was adopted.

Additional reasons for the choice are that the log-linear analysis of contingency data:

1. Provides for the systematic analysis of various interaction effects between variables to determine if these effects are significant.
2. Involves the fitting of models to data to provide a conceptual framework for further investigation of the observations encoded in the data.
3. Provides estimates of the parameters of the model that enable the significance and direction of the association between individual categories of variables to be investigated.

A context for discussion

To facilitate the discussion of log-linear modelling and the identification of some of its strengths and weaknesses a context is provided. A set of data that was collected from 1435 children from years 5, 6, 7 and 8 in N.S.W schools. The children were shown a number pattern (see figure 1) common in year 7 introductory algebra texts (Pegg & Redden, 1990) and then were asked three questions based on the pattern. The questions were:

1.Describe in words how to calculate the number of matches needed for a chain of squares of any length.

2. Calculate how many matches were used for 83 squares.
3. Express the answer to part (1) in mathematics symbols/algebra.

Here are some chains of squares built using matches.

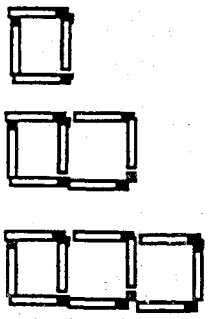

Figure 1
Number pattern Stimulus item

These three questions, together with the school year of the children became the variables of this analysis. The children's responses were coded into groups of similar responses. These groups became the categories of the variables to be investigated in the model. The frequency of each category for each variable is reported in table 1. The bold notation in parentheses will be used to refer to

the variables and categories in the remainder of this paper.

Table 1
Frequencies of response categories

School Year (Y)	Year 5	Year 6	Year 7	Year 8
	295	306	410	424

Use of Natural Language (L)	No operation described (PS)	Described one example (1EG)	Described a successive operation(SUCC)	Described a Function (FUNC)
	455	306	294	380

Calculation of an uncountable example (U)	Incorrect response (W)	Correct response(C)
	551	884

Use of Symbolic Language (S)	No Attempt (NA)	Arithmetic operations only(OS)	Incorrect use of letters(REPT)	Algebra(ALG)
	557	279	322	277

The choice of the number of categories is influenced by the decision to use log-linear analysis. The sample size should be sufficient to yield five times the number of subjects as there are cells in the contingency table and, that no more than 20% of the cells of the effects included in the model should have an expected frequency of less than five. (Norusis, 1990) . The data used here originally had 5x5x3x4 (LxSxUxY) cells which, with n=1435, contravened the ratio of cells to sample size constraint. The researcher was faced with a choice of collecting more data or collapsing cells (Kerlinger, 1986). The former is expensive in time while the latter reduces the detail available in the data. The decision was taken to collapse cells resulting in a 4x4x2x4 contingency table being the subject of the initial analysis. The expected values are reported by the SPSS platform and should be checked for values of less than five after each phase of analysis.

Towards a conceptual framework

The four variables being considered have the potential to generate a conceptual framework of considerable complexity. Figure 2 shows the fully saturated model of 15 main, two-way, three-way and four-way effects. These effects are listed in table 2 together with the partial associations for each effect. Log-linear analysis facilitates the identification of a simpler model by identifying the fewest number of effects that satisfactorily represent the data.

In the process of choosing a parsimonious model to represent the data it was necessary to decide which of the 15 effects contributed significantly to the understanding of the data. To achieve this a search is undertaken for the model with the fewest terms that provides a goodness of fit statistic that is significant at the .05 level (Gilbert, 1981). (The goodness of fit tests examines

how closely the observed frequencies fit the expected frequencies). In using the maximum likelihood estimate for the goodness of fit (G^2) statistic the following hypothesis is tested.

H: The model is a satisfactory approximation for the data.

Hence if $P<0.05$ the hypothesis is rejected and the model is considered to be ill fitting. Thus a G^2 statistic with a $P>0.05$ is required if the model is to be considered a satisfactory approximation to the data. To identify the most parsimonious model a step down procedure was used. This involved removing one effect at a time from the model. The change in the G^2 and the resulting change in the degrees of freedom were used to test the hypothesis that the effect being investigated was zero. If the observed level of significance for the change was small, the hypothesis that the effect under investigation was zero was rejected. A significance level of $P<0.05$ was accepted as a suitable standard for this comparison. The order in which effects were removed from the model was determined by the partial associations and the associated significance value. The results of this process are recorded in table 2.

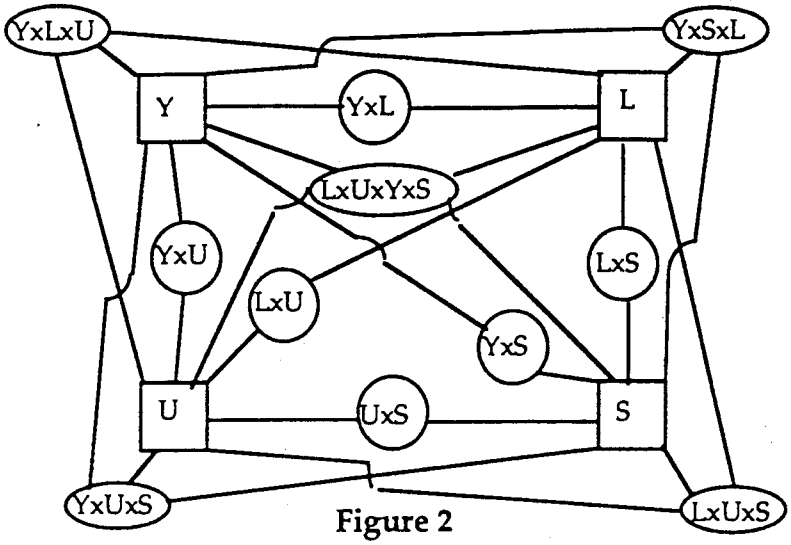

Figure 2
Saturated Model

The partial association for the effect LxUxS appears to statistically significant ($P<0.05$) however, Gilbert (Gilbert, 1981) warns that with large samples

> some relationships which are statistically significant may be of little practical significance. (p89).

This was the case with LxUxS as its removal from the model failed to make a significant difference to the G^2 value.

In addition to the above procedures, a second test was applied to determine the suitability of the resulting model . This test involved ensuring that the fit was adequate. The advice of Gilbert (1981) was accepted in setting the significance level at $P>0.05$ as an indication that the data fits the model to an adequate degree. The resulting model (figure 3), which included 5 two-way effects and the four main effects had a G^2 value of 83.39 with $P=0.498$. By comparing figure 3 with figure 2 it can be seen that a much simpler conceptual framework has been identified that satisfactorily represents the data.

Table 2
Effect names, partial associations and step down values

Effect Name	Tests of partial associations. DF	Partial Chisq	Prob	Step down values Change in G^2 on removal of this term	Change in degrees of freedom on removal of this term
YxLxUxS				16.79*	20
YxUxS	9	7.286	0.6074	7.29*	12
YxU	3	2.185	0.5348	1.02*	1
YxLxU	9	8.286	0.5056	8.60*	10
YxLxS	27	32.023	0.2312	34.99*	28
LxUxS	9	17.54	0.0409	14.70*	13
LxU	3	67.863	0		
LxS	9	683.282	0		
UxS	3	98.813	0		
YxL	9	54.769	0		
YxS	9	200.525	0		
L	3	45.644	0		
U	1	77.983	0		
S	3	136.935	0		
Y	3	38.436	0		
				* not significant at 0.05	

The systematic nature of this search assists in avoiding two possible criticisms of data analysis. The first is "data snooping" in which data is randomly investigated in the hope of finding significant relationships which had not been predicted in hypotheses developed with the guidance of a theoretical framework. The second is ignoring the possibility of interaction between variables. Using the step down procedure these interaction terms have been systematically investigated for significance.

Following the identification of a model that adequately represents the data the interaction terms that have been included in the model can be further investigated.

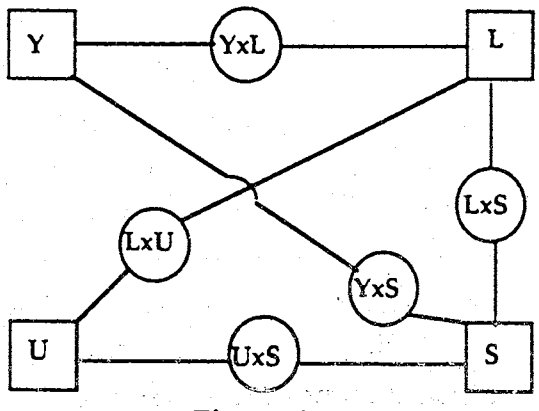

Figure 3
Parsimonious conceptual framework

The nature of the associations between variables

The inclusion of the interaction terms in the model mean that at least one of the set of possible interactions between the categories of the variables involved in the interaction are significant. The parameter estimates provide a method of further investigating the nature and level of significance of the interactions. (A suitable α value for the investigation of

interactions between categories of variables is discussed below.)

The log-linear procedure of SPSS (Norusis, 1990) produces parameter estimates (γ coefficients) for each cell. The parameter estimates provide a means of assessing the importance of the effects in determining cell frequency (Everitt, 1977). These lambda coefficients can be expressed as standard scores ($z=\gamma/se$) and used to measure statistically the strength and direction of association between categories of components. For example, if a criterion of $p<.05$ was assumed the table of γ/se values would be scanned for $|z|$ values exceeding 1.96; such values would indicate a significant association between categories of the components represented by the rows and columns of the table. In this context a positive association between categories means that a subject scoring high on one component will tend to score high on the other component, and a negative association means that a subject scoring high on one component will tend to score low on the other.

Before examining the γ/se values in more detail, it is necessary to consider an appropriate α level for the full set of tests. Because the tests are non-independent, a very conservative α level needs to be set to guard against a rapidly escalating type 1 error rate. A suitable pair-wise error rate can be calculated using the following formula which relates the family-wise error rate (α_{FW}) to the number of comparisons (C) and the pair-wise error rate (α_{PW}).

$$\alpha_{FW} = C\alpha_{PW}$$

Thus to achieve a family-wise error rate of 0.05 when making the 64 comparisons with this set of data, a pair wise α of 0.0005 (Z=3.291) would seem adequate protection.

As an example of this analysis a detailed consideration of the association between the variables natural language (L) and use of symbols (S) is undertaken. The inclusion of the LxS term in the model implies that at least one category of L has a significant association with at least one category of S. The contingency table for the LxS effect is presented as table 3.

Table 3
Contingency Table for LxS

Language (B)	Symbols (S)				
	PS	OS	REPT	ALG	Total
PS	374	41	26	14	455
1EG	93	149	56	8	306
SUCC	53	30	181	30	294
FUNC	37	59	59	225	380
Totals:	557	279	322	277	1435

The major feature of the table is the high frequencies of the cells on the diagonal from top left to bottom right and the relatively low frequencies of the cells furthest away from this diagonal. The question that needs to be asked is, are these apparent associations significant or could they have occurred by chance. To assist in answering this question SPSS provides the cell parameter estimates (γ values) in the form detailed in table 4. It can be seen that only nine parameters are given for the 16 cells in the LxS effect. The remaining cells can be calculated using the fact that the sum of each row and column is zero. Hence a complete set of parameter estimates are available as in table 5.

Before Z scores can be calculated for the seven additional parameter estimates the corresponding standard errors are required. SPSS does not provide these standard errors nor does the supporting documentation provide the algorithm to facilitate their calculation. Graybill (1961) provides a suitable algorithm under the heading of "Linear Functions of Normal Estimates" (p56).

Table 4
SPSS reporting of Parameter estimates

L BY S

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
14	1.3578703800	.09627	14.10427	1.16917	1.54657
15	.0081235182	.11331	.07169	-.21396	.23021
16	-.4164540327	.11112	-3.74765	-.63426	-.19865
17	-.2871713518	.13094	-2.19308	-.54382	-.03052
18	.8290183216	.11322	7.32228	.60711	1.05093
19	-.3841376924	.13391	-2.86861	-.64660	-.12167
20	-.6773409712	.13678	-4.95198	-.94543	-.40925
21	.0892527742	.11989	.74443	-.14574	.32424
22	.9930344714	.09409	10.55457	.80863	1.17744

Table 5
Parameter Estimates for LxS

Language	Symbols (S)			
(B)	NA	OS	REPT	ALG
PS	1.358*	0.008	-0.416*	-0.950*
1EG	-0.287	0.829*	-0.384	-0.158
SUCC	-0.677*	0.089	0.993*	-0.405*
FUNC	-0.393	-0.926*	-0.192	1.512*
* Significant at P<.0005				

Table 6
Z Values for L x S (γ /SE)

Language	Symbols (S)			
(B)	NA	OS	REPT	ALG
PS	14.104	0.072	-3.748	-8.162
1EG	-2.193	7.322	-2.869	-1.353
SUCC	-4.952	0.744	10.555	-3.787
FUNC	-2.157	-4.257	-1.295	12.905

Once calculated, the Z scores (see table 6) can be considered to be normally distributed:

> .....standardised values have, asymptotically, a standard normal distribution, and may therefore be compared with the normal deviate for any particular probability level, to obtain some idea as to the 'significance' of a particular effect. (Everitt, 1977 p98).

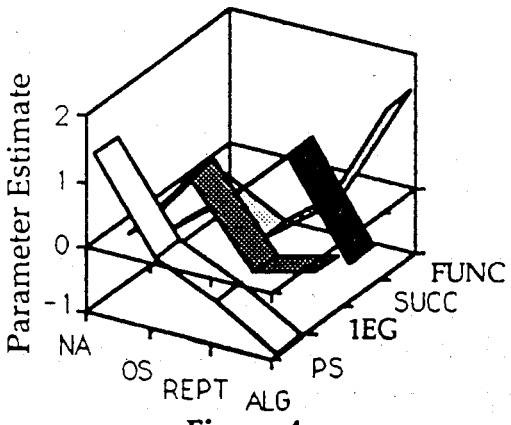

Figure 4
Response Surface

and hence the significant pairs of association can be identified. In figure 4 the ridge represents the diagonal discussed above. The pairs with a significant association all lie on this ridge, while the pairs with a significant negative association lie in the valleys.

Conclusion

In this paper a set of qualitative data in the form of coded responses to interview questions has been analysed by using the frequency of coded categories and the statistical procedures of log-linear analysis. The approach enabled a systematic investigation of the interaction effects to be undertaken. Consequently, the insignificant effects were removed from the conceptual framework, thus simplifying detailed analysis of the remaining terms. However, there was a cost in using this procedure in that some restrictions were placed on the number of categories used. The need for an overview of the data, in the form of a conceptual framework, needs to be balanced against a possible loss of detail. By focusing on the frequencies of coded sets of data the richness of the individual protocols is lost. By focusing on the individuals protocols the ability to generalise the findings is reduced. Ideally some balance between these alternatives is required. Log-linear analysis seems to have considerable potential to contribute to this balance.

References

Everitt, B. S. (1977). The Analysis of Contingency Tables. London: Chapman and Hall.

Gilbert, G. N. (1981). Modelling Society. London: George Allen and Urwin.

Graybill, F. A. (1961). An Introduction to Linear Statistical Models. New York: McGraw-Hill.

Haberman, S. J. (1978). Analysis of Qualitative Data. New York: Academic Press.

Kerlinger, F. N. (1986). Foundations of Behavioral Research (3rd ed.). Orlando: Holt, Rinehart and Winston.

Norusis, M. J. (1990). SPSS Advanced Statistics User's Guide. Chicago: SPSS Inc.

Pegg, J. E., & Redden, E. (1990). Procedures for, and experiences in, introducing algebra in New South Wales. Mathematics Teacher, 84(5), 386-391.

Tabachnick, B. G., & Fidell, L. S. (1989). Using Multivariate Statistics (2nd ed.). New York: Harper Row.