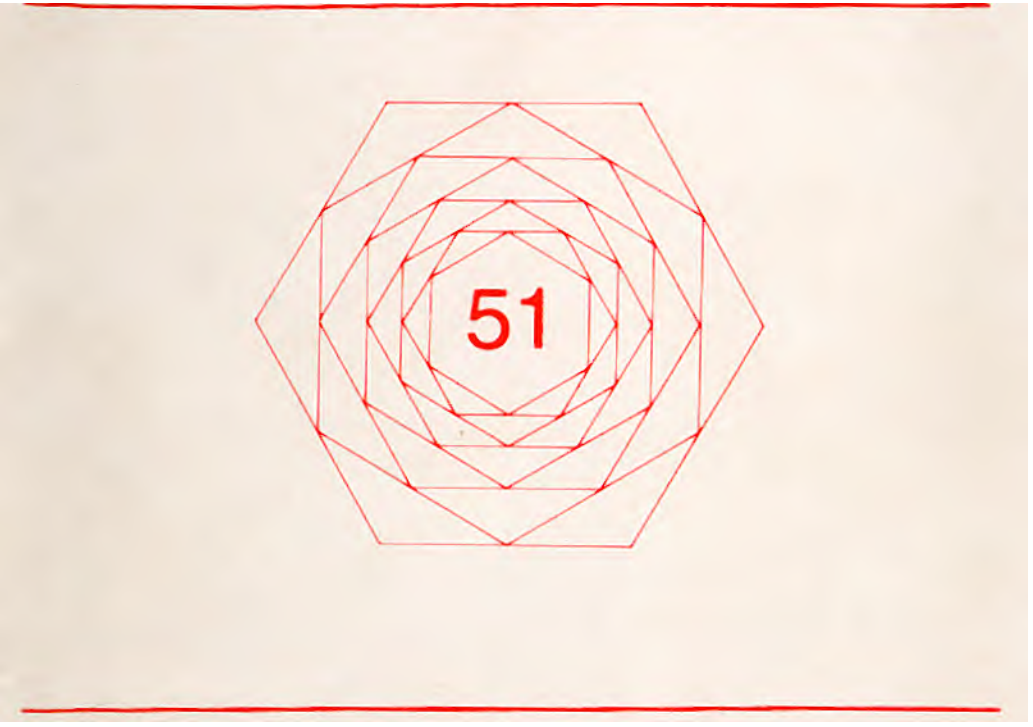
The Statistical Analysis of Contingency Table Designs



Dr. LG. O'Brien
Special Needs Information Research Unit
Department of Geography
The University of Newcastle Upon Tyne
Newcastle Upon Tyne
NE1 7RU

ISSN 0306-6142
ISBN 1 872464 01 7

Listing of Catmogs in print

CATMOGS (Concepts and Techniques in Modern Geography) are edited by the Quantitative Methods Study Group of the Institute of British Geographers. These guides are both for the teacher, yet cheap enough for students as the basis of classwork. Each CATMOG is written by an author currently working with the technique or concept he describes.

For details of membership of the Study Group, write to the Institute of British Geographers

1: Collins, Introduction to Markov chain analysis. 3.00
2: Taylor, Distance decay in spatial interactions. 3.00
3: Clark, Understanding canonical correlation analysis. 3.00
4: Openshaw, Some theoretical and applied aspects of spatial interaction shopping models. (fiche only) 3.00
5: Unwin, An introduction to trend surface analysis. 3.00
6: Johnston, Classification in geography. 3.00
7: Goddard & Kirby, An introduction to factor analysis. 3.00
8: Daultrey, Principal components analysis. 3.50
9: Davidson, Causal inferences from dichotomous variables. 3.00
10: Wrigley, Introduction to the use of logit models in geography. 3.00
11: Hay, Linear programming: elementary geographical applications of the transportation problem. 3.00
12: Thomas, An introduction to quadrat analysis (2nd ed.). 3.00
13: Thrift, An introduction to time geography. 3.00
14: Tinkler, An introduction to graph theoretical methods in geography. 3.50
15: Ferguson, Linear regression in geography. 3.00
16: Wrigley, Probability surface mapping. An introduction with examples and FORTRAN programs. (fiche only) 3.00
17: Dixon & Leach, Sampling methods for geographical research. 3.00
18: Dixon & Leach, Questionnaires and interviews in geographical research. 3.50
19: Gardiner & Gardiner, Analysis of frequency distribution .(fiche only) 3.00
20: Silk, Analysis of covarience and comparison of regression lines. 3.00
21: Todd, An introduction to the use of simultaneous-equation regression analysis in geography. 3.00
22: Pong-wai Lai, Transfer function modelling: relationship between time series variables. 3.00
23: Richards, Stochastic processes in one dimensional series: an introduction. 3.50
24: Killen, Linear programming: the Simplex method with geo-graphical applications. 3.00
25: Gaile & Burt, Directional statistics. 3.00
26: Rich, Potential models in human geography 3.00
27: Pringle, Causal modelling: the Simon-Blalock approach. 3.00
28: Bennett, Statistical forecasting. 3.00
29: Dewdney, The British census. 3.50

(continued inside back cover)

Table of Contents

1. Introduction.

1.1 Definition.

Contingency tables are data tables which are created whenever categorical data are cross-classified. This occurs frequently in the collection and presentation of social survey data based on interviews, questionnaires and secondary sources (Moser and Kalton 1971, Dixon and Leach 1978). As a result, most geographers and social scientists are likely to meet this data type regularly.

Categorical data are generally thought to consist of two distinct types of measurement (after Stevens 1946):

(1) nominal measurements: simple counts, labels and names (for example, 'male', `female'; `British', 'Irish', 'American'),

(2) ordinal measurements: counts, labels and names which exhibit a qualitative relationship or rank order, (for example, 'lower class', 'middle class', 'upper class'; 'good', `indifferent', `bad').

In cross-classifying these, three distinct types of contingency table may be created (Table 1):

(1) fully nominal tables created by the cross-classification of two or more nominal variables (Table la),

(2) mixed contingency tables created by the cross-classification of nominal and ordinal variables (Table lb),

(3) fully ordinal tables created by the cross-classification of two or more ordinal variables (Table lc).

In analysing these tables it is assumed that the patterns displayed by the data may be described by numerical models, so-called log-linear' models, which are similar in many important ways to the linear regression model (Ferguson 1977). Two points of similarity should be noted: (a) log-linear models reproduce the structure of the contingency table as a series of parameters which are both linear and additive (as in regression), and (b) the descriptive information resulting from them may be compared with theoretical baselines provided by probability distributions such as chi-square. Log-linear models may thus be used to assess hypotheses about the structure of contingency tables and to test these for significance. Because of this, a contingency table can be considered to be the observed classification of a probability sample of observations, rather than merely as a static data display.

TABLE 1: Some illustrative contingency tables

(1a) The fully nominal design: Behavioural responses to crime

	ETHNIC GROUP			
	W. Indian	Asian	White	Total
Yes	41	129	182	352
No	51	53	75	179
Total	92	182	257	531

source: based on Table 1 of Smith (1984).

NOTE:
ACTION TAKEN refers to defensive measures taken by households to avoid victimisation

(1 b) The 'mixed' design: Bystander response in a medical emergency

		AMBULANCE CALLED		
		Yes	No	Total
CITY SIZE (000)	20-40	80	118	198
	5-17	11	152	163
	<5	3	90	93
Total		94	360	454

source: based on Table 3 of Brodsky and Hakkert (1985)

(1c) The fully ordinal design: Conservative Party allegiance

		SOCIAL CLASS			
		Upper middle	Lower middle	Working	Total
VOTER AGE	Old	67	71	112	250
	Middle age	97	87	129	313
	Young	14	22	32	68
	Total	178	180	273	631

source: based on Table 4.1 of Payne (1977)

1.2 Analysing contingency tables.

In spite of the predominance of this type of data in social studies (and in many types of environmental study too), the techniques available to analyse contingency tables have remained rudimentary until recently. The traditional techniques fall into two general categories (Reynolds 1977, Blalock 1979, Dixon 1981):

(1) tests of independence,

(2) tests of association (or interaction) and agreement.

TABLE 2: Some traditional techniques for analysing contingency tables,

Chi-square
Yates' corrected chi-square
Phi-square
Contingency coefficient
Cramer's V
Yule's Q and Y
Cross-product ratio
Kendall's tau
Stuart's tau
Goodman and Kruskal's gamma
Somer's D
Goodman and Kruskal's lamda
Cohen's kappa
Hildebrand, Laing and Rosenthal's del

Some examples of the techniques suitable for certain types of problem are set out in Table 2. Further details of these may be found in the three references just given and in the classic series of papers on categorical data analysis by Goodman and Kruskal (1954, 1959, 1963, 1972). Though frequently valuable, many of these procedures possess drawbacks, and with the arrival of a new approach based on the use of `log-linear' models, they have been largely superseded.

The advantages of the log-linear model as a means of analysing contingency tables may be set out as follows:

(1) it provides a more comprehensive method of describing relationships in contingency tables than is available using traditional techniques, given that many of these are limited to tables created by no more than two variables,

(2) it allows the information in a table to be described in a linear model format similar to that used for linear regression,

(3) its parameters may be estimated and tested for significance using robust, well-respected techniques such as maximum likelihood and the log-likelihood ratio statistic,

(4) it is easily calibrated and specified in popular computer packages such as GLIM, BMDP, SAS.

The literature on the log-linear model is enormous, covering developments in many different disciplines, including applied statistics, sociology, public health, political science. The main technical achievements which led to its development are set out in Birch (1963). Detailed summaries are provided in, among others, Bishop, Fienberg and Holland (1975), Fienberg (1980), Upton (1978), Haberman (1974, 1978, 1979), and Freeman (1987). In geography, textbook descriptions of the log-linear model may be found in Fingleton (1984), Wrigley (1985), and O'Brien (1989).

1.3 Prerequisites.

Most geography undergraduates encounter data presented in the form of contingency tables in methodology classes in their first or second years of study. The data type is also widespread throughout most areas of applied geography and so will be encountered in the substantive literatures of the subject. Though there are many competing techniques available to analyse contingency tables, it is the chi-square statistic which is most widely taught to geography students. Details of this procedure may be found in most introductory statistical methods texts including, among others, Ebdon (1985), Hammond and McCullagh (1978), Blalock (1979) and O'Brien (1989).

The log-linear approach incorporates the key features of the chi-square statistic as special cases of a much more general approach to contingency table analysis. The models themselves take the form of linear relationships measured in the natural logarithmic scale. Relationships between the observations in tables are reproduced as a series of additive parameters, similar in nature to the linear-in-parameters, additive structure of the linear regression model. A study of contingency table problems may also begin from a review of this model. Details of linear regression may be found in Ferguson (1977) as well as in the references cited above.

Whilst it is likely and desirable that readers will be familiar with the chi-square statistic and the linear regression model, this is not absolutely essential, because both can be treated as special cases of log-linear models (Haberman 1974). Thus an introductory statistics course could readily begin with the latter.

Some knowledge of the issues involved in measurement and classification is also desirable. Contingency tables are created whenever a series of measurements on a single observation or case are cross-classified. Such measurements and classes reflect underlying assumptions, prejudices and predilections and may be quite incorrect for the problem being studied. The ability to misinform becomes even more prevalent when the cases are geographical areas rather than people. This is because spatial classes on maps are not equivalent to cross-classifications in tables and thus need to be handled differently. Useful material on classification issues may be found in a number of other monographs in this series, for example, Johnston (1976), Johnston and Semple (1983), Openshaw (1983), Kirby (1985), and Dixon and Leach (1978, 1984).

1.4 Terminology and notation.

Before proceeding to describe how the information in contingency tables may be analysed, it is useful to introduce some of the terminology and notation which are frequently applied to them. The following are likely to be met in the literature:

(1) Two-way table: a contingency table created by the cross-classification of two categorical variables (similarly, three-way and four-way tables.) All of the displays in Table 1 are two-way contingency tables, whereas Table 12 is three-way.

(2) 2x2 contingency table: a two-way contingency table in which each variable possesses two levels, for example, Table 3a. A more general version of this is to describe a two-way table as U, where I indicates the row variable and J the column variable. Thus Table la is a 2x3 table, Table lb a 3x2 table, and Tables lc and 22 are 3x3 tables. Similarly, UK may be used to describe a three-way table, where K refers to a third variable comprising the table. Table 12 is an example of a 2x2x2 table.

(3) Cell or elementary cell: the cross-classification of two binary variables (that is, variables possessing two levels such as the sex and tranport mode variables in Table 3a) creates four possible types of combination: male car users, female car users, male non-car users, female non-car users. These combinations are the 'cells' of a contingency table. Similarly, a 2x3 table (Table la) will have six cells, and a 2x2x2 table (Table 12) will have eight.

(4) Observed cell frequencies: the numbers of individuals in the sample who are classified in each cell. For example, 137 out of 480 respondents in Table 3a are classified as male car users. This is the observed frequency of car users in the sample of 480 respondents.

(5) Expected cell frequencies: the numbers of individuals in the sample who would be classified in each of the cells if a particular hypothesis about the table were true. The calculation of these frequencies and their comparison with theoretical benchmarks is discussed in detail in this monograph.

(6) Grand total: the total sample size. For example, 531 in Table la and 454 in Table lb.

(7) The row and column marginals: the total number of observations in each of the rows and columns of the table. In Table la the row marginals are 352 and 179, and the column marginals are 92, 182 and 257.

There are a number of different ways of representing algebraically these features of a contingency table and the information they contain. The principal distinction is between the observed frequencies and the series of expected frequencies which may be produced under hypotheses. One way of differentiating them is to write the observed cell frequencies as fij, and the expected frequencies as Fij. The subscripts (i,j) refer to the i=1,...I rows and the j=,1.....1 columns of the table. By adding up all the observed or expected frequencies in a row or column one produces the observed or expected row and column marginals. These may be described by replacing the subscript over which addition has occurred by a fullstop sign (.) or an addition sign (+). Thus $f_{t+} + f_{12} = f_{t+}$ (the observed row marginal for row 1) because summation is over the two column levels. Similarly, $F_{it+} + F_{21} = F_{+1}$ (the expected column marginal for column 1) because summation is over the two row levels. As summation may take place over all U elementary cells, the grand total may be written as f++ (or as N). These ideas may be clarified by Table 3. Table 3a illustrates a simple 2x2 contingency table with row and column totals identified. Table 3b reexpresses the structure of Table 3a using the notation described above.

TABLE 3: Notation applied to a two-way contingency table of observed cell frequencies.

(3a) The observed data: Sex and mode of transport

		TRANSPORT MODE: Other	Car	Total
SEX	Male	172	137	309
	Female	124	47	171
	Total	296	184	480

Source: based on Table 5 of Davidson (1976)

(3b) The corresponding notation

		COLUMN VARIABLE (J): j=1	j=2	Marginal
ROW VARIABLE (I)	i=1	f_{11}	f_{12}	f_{1+}
	i=2	f_{21}	f_{22}	f_{2+}
	Marginal	f_{+1}	f_{+2}	f_{++}

NOTE: For a table of expected cell frequencies, replace the letter f by capital letter F.

2. Traditional analysis: The hypotheses of Independence and interdependence

The principal advice given in textbooks to geographers faced with analysing c ontingency table data such as that in Table la, is to calculate the chi-square statistic for it. Chi Square is a statistical measure which is used to assess whether the row and column variables of a two-way table are independent of each other. This assessment involves specifying a null hypothesis which states that the two variables are independent, and an alt ernative (or research) hypothesis which states that they are not. If the former is accepted, the assessment implies that the effects of both variables are indeed independent of each other. However, if it is rejected, then the effects of both variables are assumed to be associated in some way, suggesting that it is the interdependence of the two variables which gives rise to the observed pattern of observations in the table.

Independence is not an arbitrary phenomenon. It is a probability concept wh

column classification. To test this null hypothesis of independence, researchers must calculate a table of expected frequencies under the hypothesis (Table 4) for comparison with the table of observed frequencies (Table la). Under independence it can be shown that these expected cell frequencies are given from the observed row and column marginals and the grand total:

$$F_{ij} = \frac{f_{i+}}{N} \cdot \frac{f_{+j}}{N} \cdot N \qquad (1)$$

(The first two terms in the equation are actually probabilities. It is possible to express the chi-square statistic, and the log-linear models, solely in terms of these, though this adds an extra level of complexity to the discussion.)

TABLE 4: Expected cell frequencies under independence for Table la.

		ETHNIC GROUP: W. Indian	Asian	White	Total
ACTION TAKEN	Yes	61	120.6	170.4	352
	No	31	61.4	86.6	179
	Total	92	182	257	531

Having generated these expected frequencies, the differences between the observed and expected frequencies in each cell need to be calculated:

$$\text{differences} = f_{ij} - F_{ij} \quad \text{for } i=1,...I,\ j=1,...J \text{ cells.} \qquad (2)$$

Because the observed data are the product of a sample, and therefore, possess sampling variability, it is perfectly possible that these observed differences are random or chance variations rather than a true reflection of the relationship between the row and column variables. To assess this, the differences are used to compute a measure which may be compared with a theoretical benchmark. This measure is the chi-square statistic:

$$X^2 = \text{sum} \frac{(f_{ij} - F_{ij})^2}{F_{ij}} \qquad (3)$$

where sum reflects the addition of the standardised squared differences over the IJ cells of the table.

The reason for the use of this measure is that it can be shown that X^2 is distributed according to the chi-square distribution, a continuous theoretical probability distribution, for given degrees of freedom. As the properties of this distribution are known, it may be used to

determine if the calculated measure may have arisen by chance (for further details of this distribution, see Hanushek and Jackson 1977, Ehrenberg 1982, O'Brien 1989).

The degrees of freedom are the number of independent items of information associated with the testing of the null hypothesis. For independence, they may be calculated directly from the dimensions of the table given that the observed data have already been used to calculate the observed row and column marginals. Accordingly, for independence, the degrees of freedom of a 2x3 table such as Table la, are:

$$DF = (I-1)(J-1) = (2-1)(3-1) = 2$$

The observed value of X^2 with two degrees of freedom may now be compared with an expected value using tables of the chi-square distribution. Table 5 gives expected values of chi-square for given degrees of freedom (set out as the rows of the table) and for different levels of significance (set out as the columns).

TABLE 5: Values of chi-square (first 10 degrees of freedom)

	Significance level			
DF	0.50	0.10	0.05	0.01
1	0.46	2.71	3.84	6.63
2	1.39	4.61	5.99	9.21
3	2.37	6.25	7.82	11.34
4	3.36	7.78	9.49	13.28
5	4.35	9.24	11.07	15.09
6	5.35	10.64	12.59	16.81
7	6.35	12.02	14.07	18.48
8	7.34	13.36	15.51	20.09
9	8.34	14.68	16.92	21.67
10	9.34	15.99	18.31	23.21

At the 0.05 significance level (a level indicating the probability of rejecting the null hypothesis when it is actually true) chi-square (DF = 2) is 5.99, which compares with the calculated value from the data of 23.5. The tabulated value is the value of chi-square which may be expected to be associated with Table la merely as the result of sampling variability. As the observed value exceeds this, we can assume that behaviour and ethnic group are indeed associated in some way, and that the null hypothesis of independence is incompatible with the data. It is thus rejected in favour of the alternative hypothesis of interdependence.

To see how this analysis may be conducted in MINITAB, see Figure 1. The output reproduced there will be used for comparison with log-linear analyses of these data using GLIM 3.77 (the most recent version of the GLIM system) to be presented in Section 3.

FIGURE 1: Chi-square analysis of Table la using MINITAB.

```
run minitab

MTB> read c1 c2 c3
DATA> 41 129 182
DATA> 51 53 75
MTB> end
MTB> chisquared c1,c2,c3

Expected counts are printed below observed counts

             C1       C2       C3    Total
1            41      129      182      352
          61.0     120.6    170.4

2            51       53       75      179
          31.0     61.4     86.6

Total        92      182      257      531

ChiSq =   6.55 +   0.58 +   0.79 +
         12.88 +   1.14 +   1.56 = 23.50

DF = 2

MTB> stop
```

NOTE: Commands prefixed by MTB> are MINITAB commands to read the data into three columns (c1-c3) and perform the analysis. The final command terminates the session. Other items are produced as output by the program.

3. Log-linear reanalysis in GLIM.

The table of expected cell frequencies under the null hypothesis of independence (Table 4) displays the following interesting features when compared with the table of observed frequencies (Table la):

(1) the overall size of the expected sample equals that of the observed sample,

(2) the expected row and column marginal totals are equal to those in the observed table.

This latter finding occurs because of the independence of the two variables and not because it has been predetermined before analysis. In fact, for a general analysis of contingency tables none of the features of the observed table are fixed in advance. The overall sample size reflects the success of sampling rather than the purposeful intention of researchers to select a given number of observations. (More complex types of table in which sampling restrictions are imposed are considered later.) In order to compare a table of expected frequencies with the observed table, it is necessary to assume that the two sets of frequencies are based on the same size sample. However, aside from this, there is no necessary assumption that any other feature of the observed table be preserved.

There are many different ways of distributing 531 observations among 6 cells. Table 6 gives a number of examples, which preserve different aspects of the observed table. As we have already argued that one specific configuration corresponds to the null hypothesis of independence, it follows that these alternative configurations must correspond to other types of hypothesis. How may these be calibrated and tested?

TABLE 6: Reallocation of given crime behaviour sample to cells.

Alternative 1

		ETHNIC GROUP			
		W. Indian	Asian	White	Total
ACTION TAKEN	Yes	88.5	88.5	88.5	265.5
	No	88.5	88.5	88.5	265.5
	Total	177	177	177	531

Alternative 2

		ETHNIC GROUP			
		W. Indian	Asian	White	Total
ACTION TAKEN	Yes	117.33	117.33	117.33	352
	No	59.67	59.67	59.67	179
	Total	177	177	177	531

Alternative 3

		ETHNIC GROUP			
		W. Indian	Asian	White	Total
ACTION TAKEN	Yes	50	50	100	200
	No	50	50	231	331
	Total	100	100	331	531

The most effective way of decomposing a contingency table into comparable items of information is to use a log-linear model. To he correct, we use a series of different log-linear models, each of which corresponds to a different hypothesis of contingency table structure. Before describing these models or considering what they may mean, it may be helpful to reanalyse Table la in GLIM (Baker and Nelder 1978, Payne 1986) to emphasise how the information in the table' may be decomposed into meaningful components. This analysis is summarised in Figure 2.

Three types of command are used here:

(1) data preparation commands,

(2) model specification commands,

(3) commands for fitting the models to data and assessing their performance.

FIGURE 2: Log-linear analysis of Table 1a in GLIM.

```
run glim

$UNITS 6$
$DATA OBS$
$READ
 ? 41 129 182 51 53 75$
$FACTOR ACT 2 ETH 3$
$CALCULATE ACT=%GL(2,3):ETH=%GL(3,1)$
$LOOK OBS ACT ETH$
      OBS   ACT   ETH
 1   41.00 1.000 1.000
 2  129.00 1.000 2.000
 3  182.00 1.000 3.000
 4   51.00 2.000 1.000
 5   53.00 2.000 2.000
 6   75.00 2.000 3.000
$YVARIABLE OBS$
$LINK LOG$
$ERROR P$
$FIT$
 scaled deviance = 161.17 at cycle  4
              d.f. =  5
$DISPLAY ER$
          estimate      s.e.  parameter
 1            4.483   0.04340          1
 scale parameter taken as  1.000

 unit  observed    fitted  residual
    1        41     88.50    -5.049
    2       129     88.50     4.305
    3       182     88.50     9.939
    4        51     88.50    -3.986
    5        53     88.50    -3.774
    6        75     88.50    -1.435
$FIT ACT$
 scaled deviance = 103.77 at cycle  4
              d.f. =  4
$DISPLAY ER$
          estimate      s.e.  parameter
 1            4.765   0.05330          1
 2          -0.6762   0.09180        ACT(2)
 scale parameter taken as  1.000

 unit  observed    fitted  residual
    1        41     88.50    -7.047
    2       129     88.50     1.077
    3       182     88.50     5.970
    4        51     88.50    -1.122
    5        53     88.50    -0.863
    6        75     88.50     1.985
$FIT ETH$
 scaled deviance = 79.753 at cycle  3
              d.f. =  3
$DISPLAY ER$
          estimate      s.e.  parameter
 1            3.829    0.1043          1
 2           0.6822    0.1278      ETH(2)
 3            1.027    0.1214      ETH(3)
 scale parameter taken as  1.000

 unit  observed    fitted  residual
    1        41     46.00    -0.737
    2       129     91.00     3.983
    3       182    128.50     4.719
    4        51     46.00     0.737
    5        53     91.00    -3.984
    6        75    128.50    -4.720
$FIT ETH+ACT$
 scaled deviance = 22.347 at cycle  3
              d.f. =  2
$DISPLAY ER$
          estimate      s.e.  parameter
 1            4.111    0.1085          1
 2           0.6822    0.1277      ETH(2)
 3            1.027    0.1213      ETH(3)
 4          -0.6762   0.09176      ACT(2)
 scale parameter taken as  1.000

 unit  observed    fitted  residual
    1        41     60.99    -2.559
    2       129    120.65     0.760
    3       182    170.37     0.891
    4        51     31.01     3.589
    5        53     61.35    -1.066
    6        75     86.63    -1.250
```

```
$FIT ETH+ACT+ACT.ETH$

 scaled deviance = 3.453e-12 at cycle  4
              d.f. =  0

$DISPLAY ERM$
                estimate      s.e.      parameter
       1         3.714      0.1562          1
       2         1.146      0.1793      ETH(2)
       3         1.490      0.1729      ETH(3)
       4        0.2183      0.2098      ACT(2)
       5        -1.108      0.2657  ETH(2).ACT(2)
       6        -1.105      0.2507  ETH(3).ACT(2)
  scale parameter taken as  1.000

 Current model:

    number of units is  6

    y-variate  OBS
    weight     *
    offset     *

    probability distribution is POISSON
        link function is LOGARITHM
        scale parameter is 1.000

    terms =  1 + ETH + ACT + ETH.ACT
$STOP$

NOTES:

(1)  Commands prefixed by $ are GLIM commands.

(2)  The operator %GL is used to generate levels of a
     categorical variable within GLIM up to the number
     set in $UNITS.  The two figures in brackets
     immediately following the operator define the
     number of levels in the factor and their
     relationship to the observed data.  Thus %GL(3,1)
     indicates that there are three levels with item 1
     at level 1, item 2 at level 2, and item 3 at level
     3.  Item 4 is set at level 1, item 5 at level 2
     and item 6 at level 3.  Conversely, the operator
     %GL(2,3) generates a binary factor with items 1 to
     3 at level 1, and items 4 to 6 at level 2.

(3)  Variables may be added or deleted from the fitted
     model using the + and - operators.  Thus if the
     current model is $FIT ETH+ACT, the reduced model
     including only ETH may be fitted either by typing
     (a)  $FIT ETH (this eliminates the current model
     and adds ETH to the grand mean effects model), or
     (b)  $FIT -ACT (this deletes ACT from the current
     model).
```


The data preparation commands - $UNITS, $DATA, $READ, $FACTOR and SCALCULATE - are used to specify th structure of the data in Table la for entry and processing in GLIM. (These commands - termed 'directives' in GLIM terminology - may be truncated, as may all other GLIM commands.) A description of their meaning is given in Table 7. In contrast with these, model specification merely requires researchers to select appropriate optionsfor the following commands:

(1) $YVARIABLE

(2) SLINK

(3) $ERROR

The first defines the response variable for the log-linear model, the second, defines how a function of that response (known as the predictable mean) is to be related to the explanatory information, the third, defines a probability process with which to assess departures of the observed from the expected values. For contingency tables such as Table la, it is usual to define the cell frequencies as the response, a logarithmic link function (SLINK LOG) and a Poisson error process ($ERR P). (For details of the Poisson probability distribution see O'Brien (1989).)

TABLE 7: GLIM commands to set up and tit log-linear models

Command	Use
$UNITS	Specifies the number of cells in the table
$DATA	Defines a variable to contain the observed frequencies
$READ	Prompts GLIM to read data from keyboard into the variable declared by $DATA
$DINPUT	Prompts GLIM to read data from an external file
$FACTOR	Defines the cross-classifying variables for the table and sets the number of levels in each
$CALCULATE	Calculates a series of indices which associate the observed data with the levels of the cross-classifying variables
$YVARIABLE	Defines which of the variables is to be the 'response' in the models
$LINK	Declares a logarithmic link function to relate the predictable part of the 'response' to the explanatory variables
$ERROR	Declares the form of the probability process for assessing residual variability for the models
$FIT	Prompts GLIM to fit a specified log-linear model to the table
$DISPLAY	Displays goodness-of-fit and other summary statistics,
$LOOK	Prompts GLIM to print selected items for inspection
$CYCLE	(And $RECYCLE) Commands to control the iterative estimation procedure.

NOTE: The $ sign is an integral part of the command but may vary depending on the installation of GLIM.

A series of log-linear models may now be fitted to the raw data using $FIT commands. The effects of these are summarised in Tables 8 and 9. Notice that as the number of terms associated with the $FIT commands increases, the values for 'scaled deviance' and the `degrees of freedom' in Table 8 decrease, and the tables of expected cell frequencies under different hypotheses (Table 9) preserve more of the characteristics of the observed table, until finally, they reproduce it exactly. (An interpretation of these tables is given in section 4.2.)

TABLE 8: Analysis of scaled deviance table for models applied to Table Ia

Model	Scaled Deviance	DF	Change in Scaled Deviance	DF
1. $FIT	161.17	5		
2. $FIT ACT	103.77	4	57.40	1
3. $FIT ETH	79.75	3	81.42	2
4. $FIT ETH+ACT	22.35	2	138.82	3
5. $FIT ETH+ACT+ ETH.ACT	0.0	0	161.17	5

NOTE: The values in the Change columns are calculated by subtracting the scaled deviance and DF values for models 2 to 5 from the values for model 1. The effect on scaled deviance of adding a term may be assessed by subtracting its scaled deviance value from that of a model which contains the same set of terms less that being investigated. Thus the effect of ETH.ACT is 22.35 (scaled deviance of model 4 minus that of model 5).

4. Interpreting the GLIM output.

4.1 Scaled deviance.

In attempting to interpret the models readers must look at the components being fitted, the `scaled deviance' value and the 'degrees of freedom'. The scaled deviance is a GLIM term which represents the amount of unexplained variability in a contingency table (rather like the residual sum of squares in regression). The largest value of scaled deviance is associated with the first of the fit commands, FIT , and sets the upper limit of the unexplained variability in this data set. Subsequent $FIT commands reduce this, until zero scaled deviance and zero degrees of freedom are reached with the final model ($FIT ACT+ETH+ETH.ACT), which reproduces the observed frequencies exactly. This is termed a 'saturated' log-linear model. The intermediary models are termed 'non-comprehensive' and 'independence' models. In these, some of the variability in the table is left unexplained. The degrees of freedom associated with these represent the number of independent items of data left

TABLE 9: Expected cell frequencies associated with GLIM analyses

Model 1: $FIT

		ETHNIC GROUP W. Indian	Asian	White	Total
ACTION TAKEN	Yes	88.5	88.5	88.5	265.5
	No	88.5	88.5	88.5	265.5
	Total	177	177	177	531

Model 2: $FIT ACT

		ETHNIC GROUP W. Indian	Asian	White	Total
ACTION TAKEN	Yes	117.33	117.33	117.33	352
	No	59.67	59.67	59.67	179
	Total	177	177	177	531

Model 3: $FIT ETH

		ETHNIC GROUP W. Indian	Asian	White	Total
ACTION TAKEN	Yes	46	91	128.5	265.5
	No	46	91	128.5	265.5
	Total	92	182	257	531

Model 4: $FIT ETH.ACT

		ETHNIC GROUP W. Indian	Asian	White	Total
ACTION TAKEN	Yes	60.99	120.65	170.37	352
	No	30.01	61.35	86.63	179
	Total	92	182	257	531

Model 5: $F1T ACT+ ETH +ETH.ACT

		ETHNIC GROUP W. Indian	Asian	White	Total
ACTION TAKEN	Yes	41	129	182	352
	No	51	53	75	179
	Total	92	182	257	531

unmodelled in the table. The values associated with the scaled deviance measure are thus bounded by upper and lower limits. The lower limit is always zero - indicating that all the variability has been explained by a model - and is most usually only achieved by fitting the saturated model. The value of the upper limit depends on the data being modelled. It may thus vary if the table is modified in some way.

Table 8 illustrates how the values for scaled deviance and degrees of freedom change as different models are fitted to the table. The columns headed 'scaled deviance' and 'DF' list the amounts of unexplained variability left in the table after each model has been fitted. The columns headed 'Change in ..' list the differences in scaled deviance and in the DF values between each model and Model 1. (Readers should note that comparisons may be made between any two log-linear models.) For the data in Table la the following features should be noted:

(1) the scaled deviance value is highest in Model 1 and declines as terms such as ACT and ETH are added,

(2) the scaled deviance value is smallest in the most complex model (Model 5),

(3) the addition of terms such as ACT and ETH reduces the number of degrees of freedom from a maximum of 5 (Model I) to 0 (Model 5),

(4) the effect of adding or deleting specific terms may be determined by looking at the change columns. For example, adding ACT to model 1 reduces scaled deviance by 57.4 and the degrees of freedom by I. This suggests that ACT accounts for about 1/3 of the total variability (scaled deviance = 161.I7) in the table,

(5) the expected cell frequencies produced under Model 4 are identical to those produced for the chi-square analysis in MINITAB given the rounding of the MINITAB output to one decimal place (compare Tables 4 and 9).

(6) The scaled deviance associated with Model 4 (22.35) is very similar to the chi-square value (23.5) produced by MINITAB (Figure 2).

4.2 Models.

It is now valuable to look a little more closely at what these various terms in the log-linear models actually mean. The log-linear model associated with the FIT command fits what is termed a 'grand mean' effect to the table. This is represented algebraically in the models and summary tables by the number I (earlier versions of GLIM use the notation %GM to represent this term). This represents some form of average expected cell frequency for the table as a whole. Notice that the expected cell frequencies under this model are identical, indicating that each of the six combinations of action and ethnic grouping are equally likely (Table 9, Model 1). Wrigley (1985), defining model terms using probabilities, refers to this as an `equiprobability' model.

The two non-comprehensive models (Models 2 and 3) add 'main effects' terms to the grand mean effect model ($FIT ACT fits the main effect of action, $FIT ETH fits the main effect

of ethnic grouping). The expected cell frequencies associated with these models differ from each other and from those produced by Model 1. Notice that the row marginals of Table la are preserved in Table 9, Model 2, and the column marginals in Table 9, Model 3. Model 2 indicates for each level of action available, the three ethnic groupings are equally likely, whereas Model 3 indicates for each ethnic group, both categories of action are equally likely.

Model 4 differs from the non-comprehensive models in that it includes both main effects terms simultaneously. The effect of this is to produce a table of expected frequencies which corresponds to the hypothesis of independence (compare Table 9, Model 4 with Table 4). In this, the grand total and all marginals of the observed table are preserved. Under this hypothesis the expected frequencies are determined as shown in section 2.

The fifth model adds the rather more complex term, ETH.ACT, to Model 4. This additional term represents an effect known as a `two-way interaction'. By adding this, Model 5 reproduces Table la in its entirety. The log-linear model being fitted here reduces scaled deviance to 0 for 0 degrees of freedom. This means that everything that can be described concerning the relationships between action and ethnic grouping has been, but at the expense of fitting as many terms as original cells. For this reason, Model 5 is the most complex log-linear model which may be fitted to Table la. It is usually termed a 'saturated' log-linear model.

The exact interpretation of these various terms or `parameter effects' (grand mean, main effects, two-way interaction) depends on how they are defined. In order to understand these definitions, it is important to understand something about `constraints' on model parameters.

4.3 Constraints.

Log-linear models are examples of what are termed `over-parameterised' models. This means that they contain more parameters than independent items of data. To illustrate this consider the examples in Figure 2 following the various $DISPLAY commands. After fitting Model 5, (1+ETH+ACT+ETH.ACT), GLIM generates the following 6 parameter estimates:

1	3.714
ETH(2)	1.146
ETH(3)	1.490
ACT(2)	0.2183
ETH(2).ACT(2)	−1.108
ETH(3).ACT(2)	−1.105

The addition of these six effects to $FIT is what reduces the degrees of freedom to 0. The figures in brackets refer to the level of the variable concerned. Thus ACT(2) refers to level 2 of the action variable, the behavioural response of NO. As these are the only parameters to have been produced by the saturated model, what has happened to the parameter which refers to the first level, ACT(1), the behavioural response of YES? By a similar argument, what has happened to ETH(1) and the interactions between ETH and ACT at level 1 of both?

The problem is that with six items of observed data, the cell frequencies, it is not possible to provide estimates for all the parameters which may be defined for the table. What is possible is to provide estimates for some of the parameters within each of the terms (for example, to calculate ACT(2) within the ACT main effect), but to arrange the definition of these parameters so that the omitted parameters may be inferred. This involves constraining the parameters in each term in some way so that some aspect of each term is used as a benchmark against which the parameters are defined. Such constraints are a necessary evil in that they are irrelevant for many practical purposes, for example, in the calculation of the expected cell frequencies and scaled deviance measures, but they are needed if estimates of individual parameters are required.

Two forms of constraints are most likely to be met in the literature:

(1) centre-weighted constraints

(2) corner-weighted constraints.

In the former, the terms are defined in such a way that the parameters within them are assumed to sum to zero. Thus for the main effect of action, ACT(1)+ACT(2) equals zero. Similarly, the sum of ETH(1)+ETH(2)+ETH(3) is zero. Wrigley (1985 p162) shows that this form of coding can also be applied to the interaction term. The interpretation of the parameters usually given as a result of this coding is as follows:

Main effects:	the difference in expected cell frequency of being at level i(j) of the row (column) variable rather than the overall mean,
Interactions:	the difference in expected cell value of being at level i of the row variable and j of the column variable rather than the mean.

This form of constraint system is found in computer programs such as ECTA (Fay and Goodman 1975) and BMDP (Dixon 1981), and is implicitly assumed in most of the major textbooks written on categorical data analysis (as in, for example, Bishop, Fienberg and Holland 1975). It is also the type of constraint coding used with analysis of variance models (see Silk 1981 for details) and has led to the interpretation of many of the characteristics of contingency tables in analysis of variance terms, even though the models are formally different (see Collett 1979).

The corner-weighting constraints system uses a different logic. In this, one parameter in each term is treated as a baseline against which the remaining parameters are compared. As this baseline parameter corresponds to the effects operating in a specific cell, it is feasible to consider that cell as the benchmark in comparison rather than the average of all the cells. Any of the parameters in a term can be used as the baseline, but it is most likely that either - the first or the last will generally be used. GLIM uses this form of constraint system, which is also termed `aliasing', and sets the first parameter in each term to zero. The interpretation of the effects using this system is thus:

Main effects:	the difference in expected cell frequency of being at level i (or j) of the row (column) variable rather than in the baseline cell (where i and j are not equal to 1),
Interactions:	the difference in expected cell frequency of being at level i of the row variable and level j of the column variable instead of the baseline cell (where i and j are not equal to 1).

As the baseline of each term is ACT(1) and ETH(1), the parameters generated by GLIM are interpreted as representing differences in expected cell frequencies between these and the other cells. Thus ETH(3) represents the difference in expected cell frequencies associated with the main effect of ETH of being Whit rather than West Indian. Similarly, ETH(3).ACT(2) represents the difference in expected cell frequencies associated with the two-way interaction of being an inactive White rather than an active West Indian.

From the point of view of interpretation it is important that readers realise that the values of these parameter estimates depend on the system of constraints used. Thus the analysis of Table la using ECTA or BMDP would yield different estimates to the six listed above. However, 'estimable functions' of these parameter estimates - for example, the expected cell frequencies - do not depend on the system of constraints used. They will therefore be identical regardless of which program is used. The information on scaled deviance and the degrees of freedom are also unaffected by the choice of constraints.

4.4 Estimation

The estimates of the parameters and the expected cell frequencies may be obtained in a variety of ways. The two most suitable procedures are:

(1) maximum likelihood

(2) weighted least squares.

The logic of both lies outside the scope of this monograph; details may be found in Wrigley (1985), Pickles (1985), and O'Brien (1989).

The procedure used in GLIM is maximum likelihood generated by an algorithm based on iterative weighted least squares. This procedure works by using the observed cell frequencies as initial guesses of the expected cell frequencies (hence the use of $YVARIABLE OBS in Figure 2), calculating a maximum likelihood estimate for the terms and cell frequencies based on this guess, and checking to see if any improvement is possible beyond this. By successive approximation (iteration) the trial values are improved until no further improvement is possible. At this point the algorithm is said to have converged and the estimates produced are termed the 'maximum likelihood' estimates for the fitted log-linear model. The number of iterations required to achieve convergence is displayed by GLIM as the term CYCLE. The output produced in Figure 2 indicates that four successive approximations were required to produce maximum likelihood estimates for models 1, 2 and

5, and three for models 3 and 4. (The GLIM manual should be consulted to refer to problems of non-convergence, divergence and saw-toothing - the latter referring to the tendency for the scaled deviance values to rise and fall in successive iterations.)

An alternative procedure which also yields maximum likelihood parameter estimates is termed the iterative proportional fitting algorithm, IPF (Deming and Stephan 1940). This is used in ECTA (Fay and Goodman 1975) and in C-TAB (Haberman 1973). In this procedure, the row and column marginal totals of the raw data table are used to condition the cell estimates of the tables of expected frequencies under different hypotheses. To illustrate this for the hypothesis of independence, consider Table 10. To start, the cell estimates are initially set at 1 and row and column totals are calculated (Table 10a). Then, the estimated cell frequencies are multiplied by the appropriate row (or column) marginals of the raw data, and the products divided by the row (or column) marginals of the estimated table. This leads to a revised set of cell estimates which preserve the observed row marginals (Table 10b). Notice, that Table 10b produces estimates which are identical to those produced by Model 2 in Table 9.

TABLE 10: Illustration of the iterative proportional fitting approach

10a: Initial Guess

		ETHNIC GROUP			
		W. Indian	Asian	White	Total
ACTION TAKEN	Yes	1	1	1	3
	No	1	1	1	3
	Total	2	2	2	6

10b: Step 1

		ETHNIC GROUP			
		W. Indian	Asian	White	Total
ACTION TAKEN	Yes	117.3	117.3	117.3	352
	No	59.7	59.7	59.7	179
	Total	177	177	177	531

NOTE: For cell (1,1): 117.3 = (1 x 352) / 3

10c: Step2

		ETHNIC GROUP			
		W. Indian	Asian	White	Total
ACTION TAKEN	Yes	61	120.6	170.4	352
	No	31	61.4	86.6	179
	Total	92	182	257	531

NOTE: For cell (1,1): 61 = (117.3 x 92) / 177

This is step 1 of the procedure. In step 2, the cell estimates in Table 10b are multiplied by the appropriate column (or row) marginals of the raw data, and the products divided by the estimated column (or row) marginals from Table 10b. This yields the estimates in Table 10c, which are identical to those produced by Model 4 in Table 9. As both sets of observed marginals are now preserved and have been used in the calculation of the expected cell frequencies, Table 10c corresponds to the hypothesis of independence.

A third method of estimation which may be used is weighted least squares. This approach is a modified, generalised form of the ordinary least squares approach frequently used with classical regression and differs from the procedure used in GLIM by being non-iterative. (Details are to be found in, among others, Wrigley 1985, Pindyck and Rubinfeld 1976.) The computer program, GENCAT (Landis et al I976) uses this approach. For a null hypothesis/model which is correct the three approaches produce identical estimates. The choice of method is thus a relatively minor consideration and, in practice, is likely to be dictated by the availability of software.

4.5 Testing.

In GLIM the scaled deviance measure associated with log-linear models corresponds to a measure of goodness-of-fit known as the maximum log-likelihood ratio statistic G^2. This may also be written as

$$-2(\log(l_c) - \log(l_{max})) \qquad (5)$$

where $\log(l_c)$ refers to the maximum log-likelihood of the model currently fitted to the data, and $\log(l_{max})$ refers to the maximum log-likelihood associated with the saturated model. It can be shown that this measure is approximately distributed as chi-square with the degrees of freedom being equal to the difference in the number of parameters between the two models.

The maximum log-likelihood of a model respresents the amount of variability unexplained by it. Large values indicate a model which fits the data poorly, whereas smaller values indicate a better fit. However, because the measure is not standardised to lie in a given interval (unlike the coefficient of determination which ranges from 0 to 1), to check the fit of a model we need to compare its observed 0^2 value against a tabulated chi-square value for similar degrees of freedom.

We have already seen how these degrees of freedom may be calculated directly for the independence model. Though degrees of freedom are produced as output with most computer programs the logic behind their calculation needs to be understood, as it is by no means always straightforward (Table **11).** The number of degrees of freedom under saturation is equal to 0 because every independent parameter which can be fitted to the table actually is. Under the null model, Model 1, the number of degrees of freedom is at its maximum, being equal to the number of cells in the observed table minus 1 (for the grand mean effect). As each main effect is constrained, there is one independent parameter less than the number of levels in each variable. Thus for ACT, with two levels, there is only one independent

TABLE 11: Calculation of degrees of freedom for an IJ table

model	degrees of freedom
saturation	IJ-(1+(J-1)+(I-1)+(I-1)(J-1))
independence	IJ-(1+(I-1)+(J-1))
non-comprehensive	IJ-(1+(I-1)) or IJ-(1+(J-1))
grand mean	IJ-1

NOTE: Using these relationships it is relatively easy to calculate the degrees of freedom for many other types of table. I refers to the number of levels in the row variable, J to the number of levels in the column variable.

parameter, whereas for ETH with three levels, there are two. The number of independent parameters in the interaction term may be worked out by extension.

Using this information the level of scaled deviance associated with each of the models fitted to Table la (see the scaled deviance and degrees of freedom columns in Table 8) may be assessed for significance by comparing them with values expcted from the chi-square distribution (Table 5). At the 0.05 significance level, the critical values for 2, 3, 4 and 5 degrees of freedom are 5.99, 7.82, 9.49 and 11.07 respectively. As the observed values for scaled deviance for these degrees of freedom exceed these in every case, we can assume that the expected frequencies from Models 1 to 4 are significantly different from those observed at the 0.05 significance level. These models are thus rejected in favour of Model 5 as the most appropriate description of the observed data.

An alternative way of testing each model is to use the information in the CHANGE columns in Table 8. These figures correspond to the effect of adding specific terms to the previous model and so reflect variation explained by these terms. As the values for the change in scaled deviance are greater than the critical values expected at the 0.05 significance level, their associated terms must be included in the model describing the data. An interpretation using either the unexplained variability of the scaled deviance information or the explained variability of the change in scaled deviance information results in the same conclusion: the saturated model is the only alternative to reproduce the patterns in the observed data satisfactorily.

However, in certain marginal cases, the results may not be the same, with one approach suggesting one description of the observed data, and the second another. In these circumstances, it is thought to be statistically more appropriate to emphasise the differences in scaled deviance values between models rather than the absolute values associated with any specific model. This is because the approximation of the scaled deviance measure to the theoretical chi-square probability distribution expressed in terms of differences is better than its alternative (Payne 1986).

5. Multiway contingency tables.

5.1 Multiway interactions.

The failure of the non-comprehensive main effects and independence models to describe Table la adequately merely confirms the evidence provided by the chi-square test. Given the need to look for the strength of interactions between ethnic grouping and action, further techniques could be used to assess, for example, the odds that whites are more likely than non-whites to take action in the face of perceived crime. Some of these procedures are outlined in O'Brien (1989). However, in tables which are more complex than Table la, saturation is much less likely to be the most satisfactory description of the data.

Table 12 is an example of a multiway contingency table: a contingency table created by the cross-classification of more than two categorical variables. Three binary categorical variables have been used here: S, for soil type, M, for moisture type, and T, for treatment type. The data come from a study conducted by Day et al (1986) on the effect of forest litter on the revegetation of coal mine soil at Black Mesa mine, Kayenta, Arizona. Because of their structure, tables such as Table 12 are termed three-way tables, 2x2x2 tables, or UK tables, where I refers to the row variable, J to the column variable, and K to the third variable. To clarify this notation, notice that Table 12 has been created by the amalgamation of two 2x2 tables. The first describes the relationship between moisture type and the treatment for undisturbed soils, the second, the same relationship for coal mine soils. Thus, I refers to moisture type, J to treatment type, and K to soil type. Marginals and expected cell frequencies under different hypotheses may be calculated for this design in the usual way.

TABLE 12: A multiway contingency table

Germination of seedlings in coal mine soil, Black Mesa Mine, Arizona.

		SOIL TYPE			
		Undisturbed Soil		Coal mine soil	
		TREATMENT		TREATMENT	
		NFL	FL	NFL	FL
MOISTURE TYPE	NR	4	26	4	14
	NR+I	28	52	14	26

source: based on Table 1 of Day et al (1986)

Note: NFL: No forest litter
FL: Forest litter
NR: Natural rain
NR+I: Natural rain plus irrigation

Log-linear models are of particular value for the analysis of multiway tables such as Table 12. This is because the approach outlined in sections 3 and 4 may be extended to accommodate the effects of the third variable. All that is needed is the addition of a three-way

interaction effect, and a new series of two-way interactions and main effects to represent the presence of the third variable. This represents a considerable improvement on traditional practices, which frequently required the multiway table to be decomposed into its constituent two-way tables which were then analysed separately. As there are three distinct ways of decomposing Table 12 (that is, creating tables of soil by moisture, soil by treatment and treatment by moisture) this practice might mean that the multiway relationships between all the variables are either not described at all, or are described incorrectly. Moreover, Upton (1978) and Fingleton (1981) note that analysing a series of essentially multiway relationships as though they were two-way can lead to paradoxical results, with interactions detected in the full table being either removed altogether or reversed in sign in the two-way tables. This type of behaviour has been known for nearly forty years and is frequently termed 'Simpson's paradox', after Simpson (1951) who demonstrated some of the peculiar effects which are possible if a multiway table is analysed as a series of two-way tables.

From the point of view of methodology, it is safer to assume that relationships are potentially multiway in form even if displayed in a two-way design. This is because it is always possible to introduce a third or fourth classifying factor to generate a multiway design from the two-way. However, if on analysing a multiway design it can be shown that three-way or four-way interactions are not significant, then it is possible to collapse the multiway design to an appropriate two-way summary form. Wrigley (1985, Chapter 9) presents further details about revising the architectural form of contingency tables.

The following log-linear models may be specified for a multiway contingency table beginning with the least complex:

1. grand mean effect model
2. non-comprehensive models
3. mutual independence
4. multiple independence
5. conditional independence
6. pairwise association
7. saturation

These are written out algebraically in Table 13. Notice that though there are essentially seven types of model to be considered, this number expands to reveal I9 different varieties. This represents a considerable growth in the number of models, parameter effects and individual parameters to be considered, reflecting the fact that as the dimensions of a table increase there is a more than proportional increase in the number of ways of forming two-way and multiway interactions. As before, these models correspond to different hypotheses about the relationships in Table 12, and preserve different aspects of the structure of the observed table.

TABLE 13: Models available for Table 12

Model type	Terms included	Terms omitted
Grand mean effect	1	S+M+T+SM+ST+MT+SMT
Non-comprehensive models	1+S	M+T+SM+ST+MT+SMT
	1+M	S+T+SM+ST+MT+SMT
	1+T	S+M+SM+ST+MT+SMT
	1+S+M	T+SM+ST+MT+SMT
	1+S+T	M+SM+ST+MT+SMT
	1+M+T	S+SM+ST+MT+SMT
	1+S+M+SM	T+ST+MT+SMT
	1+S+T+ST	M+SM+MT+SMT
	1+M+T+MT	S+SM+ST+SMT
Mutual Independence	1+S+M+T	SM+ST+MT+SMT
Multiple Independence	1+S+M+T+SM	ST+MT+SMT
	1+S+M+T+ST	SM+MT+SMT
	1+S+M+T+MT	ST+SM+SMT
Conditional Independence	1+S+M+T+SM+ST	MT+SMT
	1+S+M+T+SM+MT	ST+SMT
	1+S+M+T+ST+MT	SM+SMT
Pairwise association	1+S+M+T+SM+ST+MT	SMT
Saturation	1+S+M+T+SM+ST+MT+SMT	

Note: S - soil type
M - moisture type
T - treatment type

5.2 Fitting models.

GLIM may be used in exactly the same way as before to fit log-linear models to a multiway table. Models may be fitted either from the bottom up, beginning with FIT and ending with saturation, or from the top-down beginning with the saturated model. The effects on scaled deviance of fitting models from the bottom up are summarised in Table 14, and Figure 3.

TABLE 14: Analysis of scaled deviance table for analysis of Table 12.

Model	Scaled Deviance	DF	CV	Change in Scaled Deviance	DF	CV
Grand mean	83.38	7	12.02			
Mutual independence	6.81	4	7.78	76.57	3	6.25
Pairwise association	0.49	1	2.71	6.32	3	6.25
Saturation	0.00	0	-	0.49	1	2.71

Note: CV - critical values associated with model at 0.1 significance level

FIGURE 3: Log-linear analysis of Table 12.

```
Run glim

$UNITS 8$
$DATA OBS$
$DINPUT 11$
 File name? DATA1
$FAC S 2 T 2 M 2$
$CALC S=%GL(2,2):T=%GL(2,1):M=%GL(2,4)$
$LOOK OBS S M T$
        OBS        S        M        T
  1    4.000    1.000    1.000    1.000
  2   26.000    1.000    1.000    2.000
  3    4.000    2.000    1.000    1.000
  4   14.000    2.000    1.000    2.000
  5   28.000    1.000    2.000    1.000
  6   52.000    1.000    2.000    2.000
  7   14.000    2.000    2.000    1.000
  8   26.000    2.000    2.000    2.000
$YVAR OBS$
$LINK LOG$
$ERROR P$
$FIT$

 scaled deviance = 83.383 at cycle  4
             d.f. =  7
$FIT S+T+M$

 scaled deviance = 6.8123 at cycle  3
             d.f. = 4
$FIT T+S+M+S.T+T.M+S.M$

 scaled deviance = 0.49352 at cycle  3
             d.f. = 1

$D E$
            estimate         s.e.      parameter
       1       1.567       0.3918      1
       2       1.660       0.4126      T(2)
       3     -0.4026       0.4217      S(2)
       4       1.736       0.4113      M(2)
       5     -0.1302       0.3597      T(2).S(2)
       6     -0.9966       0.4325      T(2).M(2)
       7     -0.2065       0.3620      S(2).M(2)
      scale parameter taken as  1.000

$FIT T+S+M+S.T+T.M+S.M+S.M.T$

 scaled deviance = 0.00000000 at cycle  5

             d.f. = 0

$D ER$
            estimate         s.e.      parameter
       1       1.386       0.5000      1
       2       1.872       0.5371      T(2)
       3  -4.869e-16       0.7071      S(2)
       4       1.946       0.5345      M(2)
       5     -0.6190       0.7810      T(2).S(2)
       6      -1.253       0.5860      T(2).M(2)
       7     -0.6931       0.7792      S(2).M(2)
       8      0.6190       0.8802      T(2).S(2).M(2)
      scale parameter taken as  1.000

    unit  observed     fitted    residual
       1         4      4.000       0.000
       2        26     26.000       0.000
       3         4      4.000       0.000
       4        14     14.000       0.000
       5        28     28.000       0.000
       6        52     52.000       0.000
       7        14     14.000       0.000
       8        26     26.000       0.000
$STOP$
```

NOTE: In this example, data are entered from an external data file, DATA1, on unit 11 using the *$DINPUT command*.

Because the number and type of model which may be fitted is greater than for the two-way design, a useful initial analysis is to fit the families of effects in turn to the grand mean effect model. These are (a) the mutual independence model (includes all main effects), (b) the pairwise association model (includes all two-way effects), and (c) saturation (includes every effect defined for the table). This approach thus fits four of the seven possible model types outlined in Table 13.

By comparing the observed effect on scaled deviance reduction with tables of the chi-square distribution (Table 5) a general overall description based on the change in scaled deviance may be obtained. Critical chi-square values at the 0.1 significance level are detailed in Table 14. Using these, and looking at the effect on explained variation (the columns headed 'Change in'), we see that the family of three-way effects fitted using the saturated model is not significant at the 0.1 significance level, but that the two-way family (fitted using the pairwise association model) is. This means that the three-way interaction (STM) should be omitted from the final descriptive model for the table but that the two-way interactions (SM, ST and MT) should not. However, rather than include all three of these, it is useful to check if the reduction in scaled deviance attributable to the two-way family is divided unevenly, so that it may be possible to eliminate some of these terms as well.

5.3 Choosing an appropriate model.

In a complex contingency table many different log-linear models may be suitable descriptions of the relationships in the data. As a result it is necessary to adopt some form of model assessment strategy to identify the most parsimonious model available, that is, the model which describes as much of the information in the table in as few parameters as possible. In order to find this model, a variety of searching strategies have been devised based on sequential, or stepwise, fitting procedures. Two subtly different approaches are 'screening' and Aitkin's 'simultaneous testing procedure'.

53.1 STP

The simultaneous testing procedure (STP) developed by Aitkin (1978, 1979, 1980) is presented first because it has the more coherent basis in statistical theory. An added advantage is that it is also easy to adopt for use with GLIM. The object of STP is to obtain an assessment of the significance of each family of parameter effects (that is, all two-way effects, three-way effects and so on) whilst simultaneously controlling for the parameter effects which are already present in the model. The reason for needing a simultaneous test is that stepwise procedures, such as that outlined by the rudimentary interpretation of Table 14, are susceptible to 'order of entry' effects, distortions affecting the summary measures which simply reflect the order in which terms are added to the base model. These are sometimes sufficiently severe to give misleading significance tests.

The first step in applying STP to Table 12 involves calculating a global Type I error rate (γg) for the table: a rate designed to assess the probability of excluding all interactions from the model as being of insignificant importance even though some are likely to be significant.

This error rate may be determined from the formula

$$y_g = 1 - (1 - \alpha)^{2^r - r - 1} \quad (6)$$

where α is the significance level chosen for testing and r corresponds to the number of variables used in creating the table (that is, 3). (This test corresponds to the null hypothesis that SM, ST, MT and SMT are all equal to zero.) Aitkin suggests that a value for this global error between 0.25 and 0.5 will generally be appropriate, yielding a test which is sufficiently sensitive to identify truly insignificant terms whilst minimising the elimination of significant terms. At the 95% level (= 0.05), the error rate is 0.I85. At the 90% level (= 0.1), it is 0.34. The 90% level is thus chosen for calibrating the remaining tests.

The purpose of selecting a global error rate is to determine the degree of sensitivity needed in testing the table. Having chosen this level ($\alpha = 0.1$), the next step involves using it to calculate a series of error rates (y_f, where f = 1,3) which are specific to each family of parameter effects. Because these are specific rather than global, equation (6) is modified to reflect the number of terms (t) being tested in the family:

$$y_f = 1 - (1 - \alpha)^t \quad (7)$$

The highest order family in Table 12 - the three-way interaction - is tested first. As there is only one effect in this family, the family error rate for t =1 is:

$$y_3 = 1 - (.9)^1 = 0.1$$

This error rate for the three-way family can now be used to calculate an expected level of scaled deviance associated with the family by chance. In this case there is only a single degree of freedom associated with the three-way effect, so the level of scaled deviance which can be expected at the 0.1 significance level is 2.71 (Table 5, column 2). By comparing this value with the observed value for this family (0.49 in Table 14) we see that the 3-way effect contributes less to scaled deviance reduction than expected and so may be considered to be insignificant at this error rate. This confirms the finding of section 5.2.

Having eliminated the three-way family, the next step involves calculating an error rate for the family of two-way effects. The procedure is not quite as described for the three-way family because Aitkin (1980) argues that the error rate for the two-way family and the deleted three-way family should be pooled, that is, combined to form a new two-three-way family. This combined family consists of four terms (SM, MT, ST and SMT) and four degrees of freedom (the single degree of freedom from the three-way family plus the three degrees of freedom associated with the three two-way effects). Its error rate is given as:

$$y_{2,3} = 1 - (.9)^4 = 0.34$$

Once again, an expected level of scaled deviance associated with the pooled family by chance can be calculated using Table 5. In this test, DF is four and the significance level is 0.34. As a column for this level is not tabulated, the expected level may be calculated by

interpolation, yielding a value of 5.13. By comparing this expected value with the observed effect of the pooled family (6.81 .32+0.49 in Table 14), we see that not all of these two-way effects may be eliminated.

Though this test suggests that the pooled two-three-way family is significant, it does not imply that all the terms within it are significant. We have already seen that the three-way family on its own is not significant and should be deleted. We now need to see if some of the two-way effects can also be deleted. In other words, we need to see which, if any, of the parameter effects in the two-way family are marginal to its overall importance.

To do this a series of multiple and conditional independence models is fitted and the effects on scaled deviance checked (Table 15). This shows three possible orders of fitting the two-way interactions (note that the terms ST and TS are formally equivalent.) In each of these, the change in scaled deviance on the addition of MT remains constant at 5.92, a value which is considerably larger than the change values associated with the other two-way terms. The change values associated with these differ depending on the order in which they are fitted. This suggests that ST and SM may be of marginal importance to the family. To check if they may be eliminated, terms may be removed from the bottom of the table upwards until the critical value of 5.13 is exceeded. Two distinct models are suggested. In Table 15A, SM+SMT are eliminated leaving a model containing 1+S +T+M+ST+MT, whereas in Tables 15B and 15C, SM+ST+SMT may be removed.

TABLE 15: Further analysis of Table 12

	Scaled Deviance	DF	Change in Scaled Deviance	DF
(Model A)				
Null model	83.38	7		
+ T+S+M	6.81	4	76.57	3
+ ST	6.74	3	0.07	1
+ MT	0.82	2	5.92	1
+ SM	0.49	1	0.33	1
+ 3-way interactions	0.00	0	0.49	1
(B)				
1+T+S+M	6.81	4	76.57	3
+ MT	0.89	3	5.92	1
+ SM	0.62	2	0.27	1
+ ST	0.49	1	0.13	1
+ SMT	0.00	0	0.49	1
(C)				
1+T+S+M	6.81	4	76.57	3
+ MT	0.89	3	5.92	1
+ ST	0.82	2	0.07	1
+ SM	0.49	1	0.33	1
+ SMT	0.00	0	0.49	1

(0) Standardised regression coefficients from pairwise association model:

Term	Estimate	Standard error	SRC
ST	-0.1302	0.3597	-0.4
MT	-0.9966	0.4323	-2.3
SM	-0.2065	0.3620	-0.6

This leaves a model containing 1+S+T+M+MT (scaled deviance of 0.89 for 3 degrees of freedom). The choice from these alternatives is clear. if competing models satisfy the elimination criteria of STP, then the model containing fewer parameters should be used as it is the most parsimonious representation of the observed data. The model preferred in this case is a model of multiple independence indicating that the joint variable moisture type and treatment type is independent of soil type.

The selection of the most parsimonious model becomes rather more complex in higher-order tables where the number of possible sequences in a significant family may be very large. When this occurs, many models may satisfy the STP criterion. In an effort to simplify matters, Aitkin (1980) suggests fitting the terms within the family in any order initially, and then calculating their standardised regression coefficients (SRCs). These are equivalent to asymptotic t tests for each term and may be used to rank them in order of importance. The information needed to calculate these coefficients is provided by GLIM as part of the $DISPLAY E command. This results in the parameter estimates and their standard errors being printed. The standardised regression coefficients are produced by dividing the estimates by their standard errors (Table 15D). Having calculated these, terms should be refitted in order, largest coefficient first. Applying this to Table 12, we find that MT should be fitted before the SM and ST interactions, thus providing further support for the model chosen above.

Before finishing this section it is useful to point out that in higher-order tables a further model selection stage may be carried out to see if any of the lower-order relatives of terms in the preferred model can be eliminated. In the example described, this would mean checking to see if the main effect of soil, S, could be eliminated. Aitkin (1980) notes that in this case, testing should involve comparing the effect of the term in question with the critical level associated with a pooled error rate for all effects being fitted to the model. For a model containing 3 main effects, 3 two-way interactions and a single three-way interaction, this would involve 7 terms. In the case of Table 12, no further reduction in complexity is possible as the S main effect can be shown to be required in the final model.

53.2 Screening.

The STP procedure may be used with many computer systems which produce maximum likelihood estimates for log-linear models. However, a number of alternative procedures may also be used which are based on the idea of `screening' parameters and terms for significance. Wrigley (1985, Chapter 5.7) discusses some of the possibilities.

One of the most useful alternatives to STP, which is also available as an output option in module P4F of BMDP (Dixon 1981), is the screening strategy suggested by Brown (1976, 1981) and Benedetti and Brown (1978). The key idea of this procedure is to assess each term in a baseline log-linear model using two tests of association, a test of 'partial association' and a test of 'marginal association'. By comparing the information supplied by these tests, researchers can classify terms into one of the following three categories:

1. terms which should definitely be included in any final log-linear model,
2. terms which should definitely be excluded from any final log-linear model,
3. terms which perhaps should be included.

Members of classes 1 and 2 are terms whose significance, or lack of it, is suggested by both tests. Members of class 3 are terms whose significance is suggested by one test but not the other. They are thus components of a 'grey' area for which no hard-and-fast rules exist.

The terms `marginal' and `partial' association refer to two types of conditional test in which the contribution of a specific parameter effect is assessed by comparing the performance of a log-linear model which includes it with one which excludes it. As with STP it is possible to take as a baseline either the saturated model or the null model (or more frequently the main effects model) and exclude, or include, terms for testing as required. However, the two tests differ in the way they do this. Wrigley (1985) and the original authors describe the procedure in detail. However, to illustrate it, assume that the term to be tested is the two-way interaction between soil type and moisture type (SM in Table 13). The two tests amount to a comparison of the following four log-linear models:

marginal test of SM term: model I: 1+S+M+SM
model 2: 1+S+M
partial test of SM term: model 3: 1+S+M+T+SM+ST+MT
model 4: 1+S+M+T+ST+MT

The marginal test assesses the significance of SM by comparing the value of G^2 associated with model 2 with that from model 1, two models which differ solely by the exclusion of the SM interaction. Notice the absence of the T main effect or any interactions involving T. The partial test in comparison involves comparing the effect on G of fitting model 4 rather than model 3. Once again the SM term is removed, but model 3 contains a full set of main effects and two-way interactions. The marginal test includes only those lower-order parameters which are required to allow the SM two-way interaction to be defined: 1, S and M. The partial test includes all the two-way interactions which can be defined for the table. They may thus be visualised as representing upper and lower limits on the possible effect of the SM term on G^2.

Table 16 contains the results of a BMDP analysis of Table 12 which includes screening as an output option. The probability values highlight terms which are significant and should be included in the model, and terms which are insignificant and so should be excluded. High probability values indicate that a term probably should be excluded. Given this, it seems immediately clear from both tests' that the two-way interactions between SM and ST are insignificant, and that MT is significant. As both tests are identical when applied to main effects or the highest-order interaction, only one set of test figures have been printed for these. They show that the three-way effect is insignificant, and the main effects are significant. None of the terms fall in to the `grey area'. The same model identified by STP is identified by screening as the most parsiminious model for Table 12.

TABLE 16: Screening applied to Table 12.

EFFECT	PARTIAL ASSOCIATION DF	G^2	PROB	MARGINAL ASSOCIATION DF	G^2	PROB
T	1	28.33	0.0			
S	1	16.36	0.0001			
M	1	31.88	0.0			
ST	1	0.13	0.7181	1	0.07	0.7934
MT	1	5.99	0.0144	1	5.93	0.0149
SM	1	0.32	0.5696	1	0.26	0.6091
SMT	1	0.49	0.4823			

5.4 Residual analysis

The discussion so far has centred on the comparison of log-linear models against summary statistics. This is equivalent to comparing a regression model solely with the coefficient of determination (Ferguson L917). The main limitation of this is that it is possible for a model which appears to provide mil acceptable overall fit to be poor in parts of the table. To check this internal fit it is valuable to examine the residuals associated with the most acceptable log-linear model.

There are a variety of different types of residual which may be considered. These include simple residuals, standardi sed residuals and adjusted residuals. The first of these is calculated as the difference between the observed cell frequency in each cell of the table and the expected cell frequency •calculated for it under the best fitting log-linear model. The standardised residuals arm calculated in a similar fashion to the traditional chi-square measure, by subtracting bib expected cell frequencies from those observed, and dividing the differences by the square mot of the expected frequencies:

$$SR \simeq (f_{ijk} - \hat{f}_{ijk}) / \hat{F}_{ijk}^{0.5} \qquad (8)$$

Values of 3 or more indicate important residuals and roughly correspond to the 1% tails of the standard Normal distribution (Dobson 1983, pI0I). The adjusted residuals are calculated from these standardised residuals by dividing them by their asymptotic variance (Haberman 1974, Edwards I979). These appear to give a more precise and sensitive analysis.

GLIM automatically generates standardised residuals as part of its output from the the $DISPLAY command. However , the adjusted residuals need to be calculated separately. Defize (1980) gives the following GLIM code for their calculation:

```
$EXTRACT %VL
$CALCULATE ADJ. (00BS-%FV ) / %SQRT(%FV*(I- %FV*%VL))
```


The $EXTRACT command is used to copy components from the internalised working matrix within GLIM into vectors which may then be analysed. The actual command used here copies the estimated variances of the linear predictors into a system vector (%VL). The $CALCULATE command then generates the adjusted residuals. The system variable, %FV, contains the expected cell frequencies or 'fitted values' associated with the fitted log-linear model.

Table 17 displays the observed values and the three types of residuals associated with the best-fitting model applied to Table 12. These are all relatively small indicating that the model fits the internal structure of the table fairly well. However, a pattern does seem to exist among both the standardised and adjusted residuals with negative values at level 1 of moisture for undisturbed soils, and at level 2 of moisture for coal mine soils. The values of the residuals for treatment category 1 on the first row of the table are also very much higher than the others. The possible cause of this may be seen if separate models are fitted to the undisturbed data and the coal mine data. For the former, the saturated model is found to be most acceptable, indicating the importance of the interaction between treatment and moisture. However, for coal mine soil, a more acceptable model is independence. This suggests that the significance of the treatment-moisture interaction in the model for the full table is primarily due to the influence of the effect in the undisturbed data set.

TABLE 17: Observed values and residuals

OBS	RES	SR	ADJ
4.00	-1.238	-0.5410	-0.9420
26.00	-0.1905	-0.3722E-01	-0.7256E-01
4.00	1.238	0.7450	0.9430
14.00	0.1905	0.5126E-01	0.7257E-01
28.00	0.5000	0.9535E-01	0.1874
52.00	0.9286	0.1299	0.3021
14.00	-0.5000	-0.1313	-0.1874
26.00	-0.9285	-0.1789	-0.3021

```
Note: OBS - observed data
      RES - simple residuals
      SR - standardised residuals
      ADJ - adjusted residuals
GLIM codes $CALC RES=OBS -%FV
           $CALC SR=(OBS-%FV)/(%FV)**0.5
           $CALC ADJ=(OBS-%FV)/%SQRT(%FV*(1-%FV*%VL
                                                    ))
```

5.5 Checking for influential observations.

In addition to the analysis of residuals it is valuable to assess whether the patterns in the table are the result of unusually important observations. Tests for unusually important observations are helpful because they point towards areas in the raw data table which may behave very differently from expected. Peculiarities may be the result of the influence of

specific cells or of whole groups of cells (strata). A screening strategy may be used to check for these by identifying cells or strata whose frequencies differ from the expected frequencies under the proposed model by more than a given amount. This process is provided as an option in the P4F module of BMDP. None of the cells or strata of Table 12 appeared to be unusual at the 0.05 significance level when tested using this screening procedure.

5.6 Summary.

The material presented in this section has covered many of the most important features of log-linear modelling appied to multiway tables. The concepts of model fitting, assessment and examination have been presented, as have some of the alternative procedures available. The key points to note are the value of the linear-in-parameters format for handling multiway interactions; the multiplicity of models in the multiway table; the need for some form of assessment of the many models which may be suitable as descriptions of the data; and the need to assess the internal fit of a model, rather than merely rely on the overall measure of scaled deviance.

6. Incomplete contingency tables

In all of the preceding examples the observed contingency tables are complete, that is, they do not contain observed cell frequencies of zero. The presence of zeros can complicate the interpretation of log-linear models (a) because they can affect which parameter effects may be fitted, and (b) because they may influence the calculation of the degrees of freedom for the model. The following distinctions should be noted:

(1) sampling incompleteness: contingency tables which contain observed cell frequencies of zero due to deficiencies in sampling,

(2) structural incompleteness: contingency tables in which some of the cells are constrained prior to analysis to zero or to another fixed value. These structurally restricted cells cannot change their value under extended sampling or under different types of model.

6.1 Sampling incompleteness

Table 18 is an example of a table which contains sampling zeros. It has been created by substituting two of the positive cell frequencies in Table 12. The zero cells in Table 18 indicate the absence of the appropriate combinations of soil type, moisture type and treatment in the sample rather than theoretical restrictions. Because of their positions they do not affect the calculation of expected cell frequencies for any of the unsaturated log-linear models appropriate to the table (Table 19 presents some examples). Notice that under these hypotheses, the observed values of zero are replaced by non-zero expected cell values.

One suggestion which has been made to overcome the problems posed by zero cells, given that all the cells in a contingency table are the result of sampling, is to replace the zero values by a small positive value, say 0.5 (Goodman 1970). The effect of this on the analysis of Table

TABLE 18: A multiway contingency table with two sampling zero cells.

		SOIL TYPE			
		Undisturbed Soil		Coal mine soil	
		TREATMENT		TREATMENT	
		NFL	FL	NFL	FL
MOISTURE TYPE	NR	4	26	0	14
	NR+I	28	52	14	0

Note: NFL: No forest litter
FL: Forest litter
NR: Natural rain
NR+I: Natural rain plus irrigation

TABLE19: Expected cell frequencies for specific hypotheses applied to Table 1&

(a) FIT

		SOIL TYPE			
		Undisturbed Soil		Coal mine soil	
		TREATMENT		TREATMENT	
		NFL	FL	NFL	FL
MOISTURE TYPE	NR	17.25	17.25	17.25	17.25
	NR+I	17.25	17.25	17.25	17.25

(b) $FIT S+T+M$

		SOIL TYPE			
		Undisturbed Soil		Coal mine soil	
		TREATMENT		TREATMENT	
		NFL	FL	NFL	FL
MOISTURE TYPE	NR	11.69	23.38	2.97	5.95
	NR+I	24.98	49.95	6.35	12.71

(c) $FIT S+T+M+S.M$

		SOIL TYPE			
		Undisturbed Soil		Coal mine soil	
		TREATMENT		TREATMENT	
		NFL	FL	NFL	FL
MOISTURE TYPE	NR	3.18	31.88	0.8	8.12
	NR+I	33.48	41.45	8.52	10.55

Note: NFL: No forest litter
FL: Forest litter
NR: Natural rain
NR+I: Natural rain plus irrigation

18 is summarised in Figure 4. This presents a transcript of the analysis of Table 18 before and after the addition of 0.5 to every cell. Notice, that after the modification, GLIM has no difficulty in fitting the saturated model to the observed data. The effect on the scaled deviance statistics is to reduce their values throughout (for example, 130.88 instead of 139.84 for 7 degrees of freedom for the grand mean effect model). Without the addition of the 0.5 to the empty cells, GLIM cannot model the data adequately as the S(2).M(2) two-interaction is aliased.

Table 20 is a second example of a table containing sampling zeros. This differs from Table 18 in that the relative position of the zero cells has resulted in a marginal total of zero being produced for the first treatment category of soil type 2. This condition poses a rather different type of problem to that affecting Table 18, in that a full set of models cannot be fitted to the table, and so requires some modification in analysis. Figure 5 presents a transcript of a GLIM run attempting to fit models to Table 20 in the usual way. Expected cell frequencies for every cell are calculated for the grand mean effect and main effects models, and for every model containing two-way interactions except ST. The attempt to fit the ST interaction to the log-linear model results in considerable instability to the scaled deviance and degrees of freedom measures produced. Similar effects are produced when the three-way effect is included in the model.

TABLE 20: A second multiway contingency table with two sampling zeros

		SOIL TYPE			
		Undisturbed Soil		Coal mine soil	
		TREATMENT		TREATMENT	
		NFL	FL	NFL	FL
MOISTURE TYPE	NR	4	26	0	14
	NR+	28	52	0	26

Note: NFL: No forest litter
FL: Forest litter
NR: Natural rain
NR+I: Natural rain plus irrigation

If a contingency table has a marginal total of zero, all expected frequencies for cells comprising that marginal must equal zero. As a result, these cells do not contribute any independent information to the table and so should be deleted from the calculation of the degrees of freedom. Bishop et al (1975) and Fienberg (1980) show that the degrees of freedom for a table containing zero marginals can be estimated from the following formula:

$$DF\ (TC - EC) - (TP - EP) \qquad (9)$$

FIGURE 4: Analysis of Table 18 without and with the addition of 0.5.

```
run glim

$UNITS 8$
$DATA OBS$
$READ
  ? 4 26 0 14 28 52 14 0
$FAC S 2 T 2 M 2$
$CALC M=%GL(2,4):S=%GL(2,2):T=%GL(2,1)$
$YVAR OBS$
$ERR P$
$LINK LOG$
$FIT$

 scaled deviance = 139.84 at cycle  4
           d.f. =   7

$FIT +S+T+M$

 scaled deviance = 53.574 (change =  -86.26) at cycle  5
           d.f. =  4      (change =   -3    )
$FIT +S.T$

 scaled deviance = 49.364 (change =   -4.21) at cycle  5
           d.f. =  3      (change =   -1    )
$FIT +S.M$

 scaled deviance = 44.316 (change =  -5.048) at cycle  4
           d.f. =  2      (change =  -1     )
$FIT +M.T$

 scaled deviance = 17.583 (change =  -26.73) at cycle  5
           d.f. =  1      (change =   -1    )
$FIT +S.T.M$

 scaled deviance = 0.00009 (change =  -17.58) at cycle 10
           d.f. = 0        (change =   -1    )
    (no convergence yet)

$RECYCLE 20$
$FIT +S.T.M$

 scaled deviance = 3.999939203 (change = +3.99984837) at cycle  3
           d.f. = 1            (change = +1         )
    (change in d.f.)


$D ER$

          estimate        s.e.     parameter
     1       1.386      0.5000     1
     2     -0.6932      0.3273     S(2)
     3       1.872      0.5371     T(2)
     4       1.946      0.5345     M(2)
     5     0.07411      0.4659     S(2).T(2)
     6       0.000     aliased     S(2).M(2)
     7      -1.253      0.5860     T(2).M(2)
     8      -17.03       570.5     S(2).T(2).M(2)
     scale parameter taken as  1.000
```


Continued on Next Page

```
   unit  observed     fitted   residual
      1         4      4.000      0.000
      2        26     26.000      0.000
      3         0      2.000     -1.414
      4        14     14.000      0.000
      5        28     28.000     -0.000
      6        52     52.000      0.000
      7        14     14.000      0.000
      8         0      0.000     -0.001

$CALC OBS=OBS+0.5$

 -- change to data affects model

$FIT$

 scaled deviance = 130.88 at cycle  4
           d.f. =   7
$FIT +S+T+M$

 scaled deviance = 47.299 (change =  -83.58) at cycle  4
           d.f. =  4     (change =   -3   )
$FIT +S*T*M$

 scaled deviance =  0.000 (change =  -47.30) at cycle  9
           d.f. =  0     (change =   -4   )

$D ER$

          estimate        s.e.     parameter
     1       1.504      0.4714     1
     2      -2.197       1.491     S(2)
     3       1.773      0.5099     T(2)
     4       1.846      0.5073     M(2)
     5       1.594       1.526     S(2).T(2)
     6       1.521       1.525     S(2).M(2)
     7      -1.162      0.5604     T(2).M(2)
     8      -5.572       2.110     S(2).T(2).M(2)
    scale parameter taken as  1.000


   unit  observed     fitted   residual
      1         5     4.5000      0.000
      2        27    26.5000      0.000
      3         1     0.5000      0.000
      4        15    14.5000      0.000
      5        29    28.5000      0.000
      6        53    52.5000      0.000
      7        15    14.5000      0.000
      8         1     0.5000      0.000
$STOP$

NOTE: The * notation S*T*M = S+T+M+(S.T)+(S.M)+(M.T)+(S.M.T)
```


FIGURE 5: Analysis of a table with a zero marginal total.

```
run glim

$UNITS 8$
$DATA OBS$
$READ
  ? 4 26 0 14 28 52 0 26$
$FAC S 2 T 2 M 2$
$CALC S=%GL(2,2):M=%GL(2,4):T=%GL(2,1)$
$YVAR OBS$
$LINK LOG$
$ERROR P$
$FIT$

 scaled deviance = 142.00 at cycle  4
           d.f. =   7
$D R$

   unit  observed     fitted   residual
      1         4     18.750     -3.406
      2        26     18.750      1.674
      3         0     18.750     -4.330
      4        14     18.750     -1.097
      5        28     18.750      2.136
      6        52     18.750      7.679
      7         0     18.750     -4.330
      8        26     18.750      1.674

$FIT +S+T+M$

 scaled deviance = 29.177 (change =  -112.8) at cycle  4
           d.f. =  4     (change =    -3  )
$D R$

   unit  observed     fitted   residual
      1         4      6.884     -1.099
      2        26     25.383      0.122
      3         0      2.503     -1.582
      4        14      9.230      1.570
      5        28     16.583      2.804
      6        52     61.150     -1.170
      7         0      6.030     -2.456
      8        26     22.236      0.798
$FIT +M.T$

 scaled deviance = 22.883 (change =  -6.295) at cycle  4
           d.f. =  3     (change =  -1    )
```

```
$D R$

   unit  observed     fitted   residual
      1         4      2.933      0.623
      2        26     29.333     -0.615
      3         0      1.067     -1.033
      4        14     10.667      1.021
      5        28     20.533      1.648
      6        52     57.200     -0.688
      7         0      7.467     -2.733
      8        26     20.800      1.140
$FIT +S.M$

 scaled deviance = 22.055 (change =  -0.828) at cycle  4
           d.f. =  2      (change =  -1     )
$D R$

   unit  observed     fitted   residual
      1         4      2.727      0.771
      2        26     27.273     -0.244
      3         0      1.273     -1.128
      4        14     12.727      0.357
      5        28     21.132      1.494
      6        52     58.868     -0.895
      7         0      6.868     -2.621
      8        26     19.132      1.570
$FIT +S.T$

 scaled deviance = 0.0001 (change =  -22.05) at cycle 10
           d.f. = 1       (change =   -1    )
     (no convergence yet)

$RECYCLE 20$
$FIT +S.T$

 scaled deviance =32.306869507 (change =+32.30674362) at cycle  2
           d.f. = 2            (change = +1          )
     (change in d.f.)
$FIT +S.T.M$

 scaled deviance =  0.000 (change =  -32.31) at cycle 13
           d.f. =  1      (change =   -1    )
     (change in d.f.)

$RECYCLE 20$
$FIT +S.T.M$

 scaled deviance =32.307197571 (change =+32.30712509) at cycle  2
           d.f. = 2            (change = +1          )
     (change in d.f.)

$STOP$
```


where
TC is the total number of cells in the table,

EC is the total number of empty cells,

TP is the total number of parameters being included in the model,

EP is the total number of parameters which cannot be included because of the empty marginal.

For the model 1+S+T+M+S.T+S.M+T.M applied to Table 20 this amounts to reducing the calculated degrees of freedom from 1 to O. This is because TC is 8, EC is 2, TP is 7 (1, S, T, M, S.T, S.M, T.M) and EP is 1 (the S.T marginal contains one zero entry). Payne (I979) provides a GLIM macro (a series of commands rather like a subroutine in FORTRAN, which may be invoked from different points in the GLIM session without needing to be retyped) for calculating degrees of freedom automatically for contingency tables. This macro, CDF, may be used after the fitting of a specific model in GLIM using the $USE command. If action is required, that is, if the default calculation of DF provided by GLIM is incorrect, the macro generates the correct values. Figure 6 illustrates the use of this macro on the data in Table 20. For further details of the use of this macro, see the original article by Payne.

A third approach which is sometimes adopted to handle sampling zeros is to use prior information. Bishop et al (1975) and Wrigley (1985) suggest that the observed values of zero (and the observed cell frequencies which are the product of a specific sampling) should be replaced by a priori cell values based on previous experience or theory. The logic of this is that prior information ought to be incorporated into research wherever possible (see Ehrenberg I982 for details). However, this approach, a form of empirical or pseudo-Bayesian analysis, is fraught with difficulty, not least of which being the calculation of the prior observed cell frequencies in the first place. For a discussion of the relative merits and demerits of Bayesian approaches see Lindley (1965).

6.2 Structural incompleteness.

The second type of incompleteness which may be encountered in geographical contingency tables is termed structural incompleteness. This occurs whenever a contingency table contains cells whose expected values are not allowed to change under different hypotheses. Most frequently, structurally incomplete cells will contain zeros, and so be difficult to distinguish from sampling incomplete tables. However, any non-zero expected value which is restricted under different hypotheses will behave like a structurally incomplete zero and is usually treated accordingly.

The analysis of the structurally incomplete table depends on whether that table can be modified in some way. Table 21a illustrates the basic idea. This table contains five structurally incomplete cells. By modification (Table 21b), these cells can be gathered together to reveal a subset of the original table which is complete. Analysis using standard methods may now be applied to the complete subset

FIGURE 6: Use of macro for calculating degrees of freedom.

```
run glim

$MACRO CDF
   ? $CALC WTT=%GT(%FV,0.001)*WT$
   ? $CALC %T=%EQ(%CU(WTT),%NU)$
   ? $EXIT %T$
   ? $PRINT '*** CORRECT DF FOLLOWS ***'$
   ? $WEIGHT WTT $FIT .$
   ? $WEIGHT WT$
   ? $ENDMAC$
$UNITS 8$
$DATA OBS$
$READ
   ? 4 26 0 14 28 52 0 26$
$FAC S 2 T 2 M 2$
$CALC S=%GL(2,2):T=%GL(2,1):M=%GL(2,4)$
$YVAR OBS$
$LINK LOG$
$ERR P$
$CALC WT=1$
$WEIGHT WT$
$FIT S+T+M+S.T+S.M+T.M$

 scaled deviance = 0.0001 at cycle 10
             d.f. = 1
     (no convergence yet)
 $USE CDF$

 *** CORRECT DF FOLLOWS ***
 -- model changed

 scaled deviance = 8.845e-13 at cycle  2
             d.f. =      0     from 6 observations

 -- model changed

$STOP$
```

Unfortunately, this ability to separate a contingency table into subsets is not widespread, and geographers should expect to meet tables which cannot always be handled in this way. Table 22 presents an example in which the upper part of the table contains structural zero cells, on the assumption that wards which contained urban or mainly-urban enumeration districts in 197I would not have reverted to rural land use in 1981. In order to analyse this using log-linear models, we need to redefine the concept of independence outlined earlier. Instead of defining this with respect to all the cells of the table, it is now necessary to limit the definition to only those cells which are not subject to a restriction. This leads to what Goodman (1968) terms an hypothesis of quasi-independence. The models which may be applied to Table 22 are similarly termed quasi-log-linear models. The procedures outlined

TABLE 21: Structurally incomplete contingency tables

(a) The original table

		Variable B		
		1	2	3
Variable A	1	53	0	27
	2	0	0	0
	3	45	0	18

(s) Alter modification

		Variable B	
		1	3
Variable A	1	53	27
	3	45	18

TABLE 22: A triangular contingency table

Change in hypothetical ward classifications 1971-1981

		1981		
		Urban	Mixed	Rural
1971	Urban	235	0	0
	Mixed	57	112	0
	Rural	21	78	81

earlier for fitting and testing the effects of log-linear models in GLIM may be applied to quasi-log-linear models too. However, some adjustment needs to be made to the calculation of the degrees of freedom to reflect the structural restrictions relevant to the table. Instead of using the standard formulae for degrees of freedom, or the modifications and macro applying to sampling incomplete tables, it is necessary to apply the following:

$$DF = TC - RC - TP \qquad (10)$$

where

TC is the total number of cells in the table,

RC is the total number of restricted cells,

TABLE 23: An asymmetric contingency table

		ETHNIC GROUP			
		W. Indian	Asian	White	Total
ACTION	Yes	41	129	182	352
	Total	92	182	257	531

TP is the total number of parameters in the fitted model.

For the hypothesis of quasi-independence based on Table 22, the degrees of freedom are 9-3-3 = 3. This differs from the number of degrees of freedom which are associated with the normal independence model for this table, namely, (I-1)(J-1) = (3-1)(3-1) = 4. For further details of approaches to incomplete or otherwise complex contingency tables, see Bishop et al (1975) and Wrigley (1985).

7. The asymmetric table.

The contingency tables considered so far consist of cross-classifications in which the cell frequencies under some hypothesis are regarded as the response components in the model. However, it may be more suitable to regard one of the classifying variables as the response instead, rather like the situation in regression and the analysis of variance. In this asymmetric problem the object of analysis is to see how the classification of the response is affected by the classification of the 'explanatory' variables. Table 23 illustrates the basic idea. This table is a modification of Table la in which the action variable has been reduced to a single level corresponding to the YES response. For such a table there is a choice of methods of analysis (Wrigley 1985, chapters 4 and 6). The first is to continue to use the log-linear model on the unmodified table (that is Table la), but incorporate terms in such a way that the dependency relationship is emphasised. Alternatively it is possible to treat the contingency table as a tabular version of the logit model.

The latter approach is considered first. Unlike the situation in Table la in which the cell entries are observed frequencies, the entries in Table 23 are observed proportions. Thus for West Indians, 41 out of 92 people surveyed took action, whereas for Whites, 182 out of 257 took action. It thus seems reasonable to ask whether the propensity of respondents to take action in the face of perceived crime is related to their ethnic status. In order to do this the structure of Table 23 must be recreated within GLIM, or some other program, and the relationships between the variables specified. The following differences should be noted between the analysis of this table using log-linear models (described in Figure 2) and its logit reanalysis (described in Figure 7):

FIGURE 7: Transcription of a logit analysis of Table 23

```
run glim

$UNITS 3$
$DATA ACT TOT$
$READ 41 92 129 182 182 257$
$FAC ETH 3$
$CALC ETH=%GL(3,1)$
$LOOK ACT TOT ETH$
        ACT       TOT       ETH
  1    41.00     92.00    1.000
  2   129.00    182.00    2.000
  3   182.00    257.00    3.000
$YVAR ACT$
$ERR B TOT$
$LINK G$
$FIT$

 scaled deviance = 22.347 at cycle  3
             d.f. =  2
$FIT +ETH$

 scaled deviance = 0.000000 (change =  -22.35) at cycle
             d.f. = 0        (change =   -2   )
$D ERM$

          estimate        s.e.     parameter
     1     -0.2183      0.2098     1
     2       1.108      0.2557     ETH(2)
     3       1.105      0.2507     ETH(3)
     scale parameter taken as  1.000

   unit  observed    out of    fitted    residual
      1        41        92     41.00       0.000
      2       129       182    129.00       0.000
      3       182       257    182.00       0.000

 Current model:

   number of units is  3

   y-variate  ACT
   weight     *
   offset     *

   probability distribution is BINOMIAL
     with binomial denominator TOT
             link function is LOGIT
          scale parameter is 1.000

   terms =  1 + ETH
$STOP$
```

(1) the $UNITS command is set to 3 (because there are three ethnic groups) instead of 6, though six items of data are to be read in,

(2) two variables are defined in the $READ command, one to contain the number of persons taking action (ACT), the other to hold the number in each ethnic group (TOT),

(3) the number of persons taking action is defined as the response variable rather than the observed cell frequencies,

(4) the $ERROR command is set to B to indicate that the cell proportions are assumed to be drawn from independent binomial samples rather than Poisson samples,

(5) the TOT variable is used in the definition of the error to act as the denominator for each level of ACT,

(6) the $LINK command is set to G rather than LOG, indicating to GLIM that a logit link is to be used instead of the logarithmic link assumed in log-linear models. Apart from these differences, the structure of the program is very similar to those presented earlier.

A transcription of the GLIM analysis of Table 23 is presented in Figure 7. Two models are fitted to the table. The first is the grand mean effect model. This has a scaled deviance of 22.35 for 2 degrees of freedom, and is equivalent to a constant proportions model. In other words, the proportion of respondents taking action is estimated to be the same as the observed marginal proportion (352/531 = 0.663). The second model fits the effects of ethnic status. This contains the grand mean effect and the main effect of ethnic status, and is the saturated logit model for the table.

This model is significant at the 0.05 significance level, indicating a considerable improvement over the first model. The positive signs of the parameters indicate that Asian and White households are more likely to take action than West Indian households, the anchor cell for this analysis.

It is also possible to analyse the observed frequency data in Table la to pick out dependency relationships between action and ethnic status. This facility is made possible by the fact that many log-linear models require that certain cell marginals from the observed table are preserved in order that expected cell frequencies under specific hypotheses might be calculated. It is but a small step from preserving observed marginals in a general symmetric log-linear model to requiring that these marginals be fixed by design. However, when fitting log-linear models to contingency tables in which certain observed marginals are fixed, as in Table 23 where TOT is defined explicitly as the fixed total for each ethnic group, it is necessary to ensure that parameters associated with them are included in every log-linear model applied to the table. Thus for Table la, if ACT is assumed to be the response variable (similar to a response variable in regression) and ETH is the explanatory variable, we must automatically include the main effect of ETH in every log-linear model applied to the table. The full set of asymmetric log-linear models and their associated scaled deviance values are:

(1)	1+ETH	79.75
(2)	1+ETH+ACT	22.35
(3)	1+ETH+ACT+ETH.ACT	0.0

Model 3 is the saturated asymmetric model for the table. (These scaled deviance values may be compared with the values attained for the symmetric analysis of this table in Figure 2.)

If the contingency table contained three dimensions (ACT, ETH, and a duration variable, DUR), the full set of asymmetric log-linear models would be as follows (ACT is still assumed to be the response variable):

(1) 1+ETH+DUR+ETH.DUR
(2) Model 1 + ACT
(3) Model 1 + ACT+ACT.DUR
(4) Model I + ACT+ACT.ETH
(5) Model 1 + ACT+ACT.DUR+ACT.ETH
(6) Model 1 + ACT+ ACT.DUR+ACT.ETH+ACT.ETFI.DUR

Model 1 is the basic model reproducing the dependency structure of the table. Model 2 includes the main effect of ACT, Models 3, 4 and 5 include two-way interactions between the response and the two explanatory variables, and Model 6 includes the three-way response between the variables. What is important to note about this sequence is that the terms in Model 1 must always be present if the fixed marginals associated with ETH and DUR are to be preserved.

8. The ordinal table.

In section 1 it was noted that contingency tables could contain mixtures of ordinal and nominal variables, or be fully nominal or ordinal. The examples presented so far have limited themselves to the analysis of the fully nominal table, principally because this is the form of data most likely to be encountered and for which the statistical procedures are most fully developed. However during the 1980s a number of procedures have been developed and made available which allow the ordinal and partially-ordinal table to be modelled in linear-in-parameters form. Sonic of these models are programmable in GLIM using macros. As the coding needed to compute these is complex, corresponding to the increased complexity of the topic, this section merely seeks to outline some of the models available.

Goodman (1979, 1981a,b) considers a series of models which are applicable to contingency tables having ordered rows and/or columns. These are based on previous developments suggested by Birch (1965), Haberman (1974a) and Simon (1974). McCullagh (1980) also considers models suitable for ordinal designs. Reviews of these procedures may be found in Agresti (1984).

The procedures outlined earlier - log-linear and logit models - are not entirely appropriate for the analysis of the ordinal design because the definition of multiway interactions used by them assumes that the ordering of levels in rows and columns is arbitrary. In the ordinal design the relative position of these levels is important and so, therefore, a nominal model applied to such a table automatically fails to incorporate information which may prove to be meaningful in analysis. Agresti (1984) lists four reasons for analysing the ordinal design using measures of interaction which accommodate relative row (column) order. These are:

1. **ordinal measures have greater power for detecting important alternatives to null hypotheses, such as independence, than conventional measures,**
2. **ordinal data description is based on measures that are similar to those used to analyse continuous data, thus allowing useful conceptual links to be made between the analysis of categorical and continuous data,**
3. **ordinal measures can allow for the specification of a wider variety of models than is possible using the log-linear and logit approaches,**
4. **ordinal measures may produce models which are more parsimonious than those developed using nominal techniques.**

Many different types of procedure exist, some of which assume the presence of an underlying 'latent' continuous variable (McCullagh 1980) - equivalent in context to a tolerance distribution in biology or a utility distribution in economics - along which the observed categories may be positioned. The nature of this latent variable is questionable and McCullagh notes that it is possible to develop models for ordinal data which do not in fact make reference to the existence of such a variable. Indeed he notes:

if such a continuous underlying variable exists, interpretation of the model with reference to this scale is direct and incisive. If no such continuum exists the parameters of the models are still interpretable in terms of the categories recorded and not those which might have obtained had the defining criteria been different

(McCullagh 1980 p110)

One approach which is frequently used to accommodate ordinality is to incorporate some sort of 'score' in models, where the score represents the relative position of any ordinal category with respect to the other categories. These scores may arise from theory (for example, scores developed in stimulus-response or choice-making experiments may reflect a belief in a tolerance or a utility distribution) or be essentially ad hoc. If the latter, their form depends on a series of assumptions about how the ordinal structure should be handled. Breen (I985) notes that a variety of models for the two-dimensional 1.1 ordinal table may be described based on this concept. These include additive and multiplicative models, and models in which scores relevant to either the rows or the columns only are added. The 'basic' model in this sequence is one incorporating both row and column scores as multiplicative effects, producing an equation which is consistent with the linear-by-linear interaction model

suggested by Nelder and Wedderburn (1972), or the uniform association model suggested by **Goodman** (1979).

In addition to these models, McCullagh (1980) has suggested another series of models based on the concepts of 'proportional odds' an 'proportional hazards' (the description of these models lies outside the scope of this monograph). **Hutchison (I985) shows that** these models may also be fitted in GLIM using specially-written macros. He also notes that GLIM macros exist for the 'continuation odds' model suggested by Fienberg and Mason (1979), and the 'partial likelihood survival' model (Cox 1975, Whitehead 1980, Allison 1982). A computer package, PLUM, also exists which allows a variety of these models to be fitted to data. For further details, see McCullagh (1979).

9. Conclusions and comments.

Contingency tables are an important source of categorical data. Traditionally, this data type has been considered of limited mathematical value because of its emphasis on enumeration and ranking. Compared with the apparent sophistication of linear regression, techniques traditionally thought suitable for the analysis of categorical data are distinctly cumbersome and weak. The development of log-linear models overcomes this problem by allowing tabular arrays of categorical data to be handled by parametric models which are as powerful and sophisticated as those used in regression. Indeed, it can be shown that the popular regression model is itself a special case of the general log-linear model which is suitable for a specific type of data and analytical problem.

The advantages of log-linear models, and their equivalents for the ordinal design, over traditional approaches are that they allow a wider variety of hypotheses to be tested, they are easier to estimate and test, and they may be fitted to data without major difficulty using commercially available software. Most of these software packages contain facilities to fit log-linear models to data, either as separate modules or procedures, as in BMDP, SPSS or SAS, or as an integrated facility within a wider class of linear models, as in GLIM. Regardless of how they are treated however, the log-linear models approach provides a powerful integrating framework for the treatment of categorical data met in contingency tables; a framework that may be of particular relevance to the geographer whose work is heavily influenced by categorical data.

In spite of this, the geographer must always remember that the development of powerful tools allied to popular software in no way removes the fundamental difficulties of analysing categorical data. These concern the low-level and frequently arbitrary nature of the classifications. Few would argue that the division of humanity into male and female is arbitrary, but for many types of social, environmental or areal classification, the divisions are the result of a political or administrative process and are thus heavily influenced by the sensitivity of the generators to their data. A simple example illustrates the basic point: in

coding unemployment as a category, should one include people who are members of a training scheme?

The main problem in analysing categorical data has thus changed with the development of the log-linear model from the purely technical issue of how to search for patterns among the observations, to a consideration of how those patterns vary with classification. This means that analysts have to pay particular attention to the motivations for categorising data and the purposes for which such classifications are later put. The features presented here provide a springboard into this area of interest which brings together the important areas of numerical analysis and qualitative research.

10. References

10.1 Theory: General statistical texts and papers.

Bishop YMM, Fienberg SE and Holland PW 1975 Discrete multivariate analysis: Theory and Applications. MIT

Blalock HM 1979 Social Statistics, McGraw-Hill Kogahuska

Dobson A 1983 An Introduction to statistical modelling. Chapman and Hall.

Fienberg SE 1980 The statistical analysis of cross-classified data. MIT.

Freeman DW 1987 Applied Categorical Data Analysis. Marcel Dekker.

Haberman SJ 1974 The analysis of frequency data. University of Chicago Press, Chicago.

Haberman SJ I978 The analysis of qualitative data Vol 1 Introductory Topics. Academic Press, New York.

Haberman SJ 1979 The analysis of qualitative data Vol 2. New Developments. Academic Press, New York.

Nelder JA and Wedderburn RWW 1972 'Generalised linear models'. Journal of the Royal Statistical Society A 135, 370-384.

Payne CA 1977 The log-linear model for contingency tables. In O'Muircheartaight CA and Payne C (Eds) The analysis of survey data Vol 2. Wiley, London.

Reynolds 1977 The analysis of cross-classifications. Free Press, Glencoe.

Upton GJG 1978 The analysis of cross-tabulated data. Wiley, Chichester.

10.2 Statistical texts for the ordinal design.

Agresti A 1980 'Generalised odds ratios for ordinal data', Biometrics 36, 59-67

Agresti A 1984 Analysis of ordinal categorical data. Wiley.

Goodman LA 1968 'The analysis of cross-classified data: independence, quasi-independence and interaction in contingency tables with or without missing cells'. **Journal of the American Statistical Association** 63, 1091-II3I.

Goodman LA I970 'The multivariate analysis of qualitative data: interactions among multiple classifications'. **Journal of the American Statistical Association** 65, 226-256

Goodman LA and Kruskal W I954 'Measures of association for cross-classifications' **Journal of the American Statistical Association** 49, 732-764

Goodman LA and Kruskal W 1959 'Measures of association for cross-classifications II: further discussion and references' **Journal of the American Statistical Association** 54, 123-163

Goodman LA and Kruskal W 1963 'Measures of association for cross-classifications III: Approximate sampling theory' **Journal of the American Statistical Association** 58, 310-364

Goodman LA and Kruskal W 1972 'Measures of association for cross-classifications IV: Simplification of asymptotic variances' **Journal of the American Statistical Association** 67, 415-421

Khrishnaian PR and Yochmowitz MG 1980 'Inference and the structure of interactions in two-way classification models'. **Handbook of Statistics** I, 973-994

Roy SN and Kastenbaum MA I956 ' On the hypothesis of no interaction'in a multiway contingency table', **Annals of Mathematical Statistics** 27, 749-757

Simpson EH 1951 'The interpretation of interaction in contingency tables'. **Journal of the Royal Statistical Society B,** I3, 238-241.

Snee RD I982 Nonadditivity and a two-way classification: is it interaction or non-homogeneous variance?', **Journal of the American Statistical Association** 77, 515-519

Whittemore AS 1978 'Collapsibility of multidimensional contingency tables', **Journal of the Royal Statistical Society B** 40, 328-340.

10.5 Other statistical articles

Allison PD I982 'Discrete-time methods for the analysis of event histories'. **Sociological Methodology,** 6I-98.

Birch MW 1963 'Maximum likelihood in three-way contingency tables', **Journal of the Royal Statistical Society B** 25, 220-233

Cox DR 1972 'Regression models and life tables'. **Journal of the Royal Statistical Society B,** 34, 187-220

Cox DR 1975 'Partial likelihood'. **Biometrika** 62, 269-276

Edwards D 1979 'Analysis of residuals in two-way contingency tables'. **GLIM Newsletter** 1, 30-31

Fienberg SE and Mason W 1979 'Identification and estimation of age, period and cohort models in the analysis of discrete archival data'. **Sociological Methodology,** 1-67

Whitehead J 1980 'Fitting Cox's regression model to survival data using GLIM', **Applied Statistics** 29, 268-275.

10.6 Computer software

Baker RJ and Nelder JA 1978 **The GLIM system: Release 3.** Numerical Algorithms Group Ltd, Oxford.

Breen R 1985 'Log-multiplicative models for contingency tables using GLIM'. **GLIM Newsletter** 10, 14-19

Collett D 1979 'Review of GLIM 3 Users Manual', **Biometrics** 35, 527-528.

Defize PR 1980 'The calculation of adjusted residuals for log- linear models in GLIM', **GLIM Newsletter,** 3, 41

Deming WE and Stephan FF I940 'On a least squares adjustment of a sampled frequency table when the expected marginals are known', **Annals of Mathematical Statistics 11,** 427-444.

Dixon WJ I981 **BMDP: Biomedical computer programs.** University of California Press.

Fay RE and Goodman LA 1975 **The ECTA program: Description for users.** Department of Statistics, University of Chicago.

Haberman SJ I973 **C-TAB: Analysis of multidimensional contingency tables by log-linear models:** Users **Guide.** International Educational Service. Chicago, Illinois.

Hutchinson DA 1985 'Ordinal variable regression using the McCullagh (proportional odds) model'. **GLIM Newsletter** 9, 9-I7

Landis JR, Stanish WM, Freeman JK and Koch GG 1976 A computer program for the generalised chi-square analysis of categorical data using weighted least squares (GENCAT)'. **Computer programs in Biomedicine** 6, 196-231.

McCullagh P 1979 **PLUM: An interactive computer package for analysing ordinal data.** Department of Statistics, University of Chicago, Chicago, Illinois.

Payne C 1979 'A macro for calculating the correct degrees of freedom for a log-linear model of a contingency table', **GLIM Newsletter** 1, 32-33.

Payne C (Ed) 1986 **The GLIM system: Release 3.77.** Numerical Algorithms Group Ltd., Oxford.

10.7 Quantitative geography texts and papers

Brodsky H and Hakkert AS I985 'Bystander response in an emergency', **Transactions of the Institute of British Geographers 10,** 303-3I6

Davidson RN 1976 **Causal inferences from dichotomous variables. CATMOG** 9, Norwich.

Day AD, Ludeke KL and Thames JL 1986 'Revegetation of coal mine soil with forest litter', **Journal of Arid Environments I1,** 249-253.

Dixon CI and Leach B I978 **Sampling methods for geographical research. CATMOG** 17, Norwich.

Dixon CJ and Leach B 1984 **Survey research in underdeveloped countries. CATMOG** 39, Norwich.

Ebdon D 1985 **Statistics in Geography: A Practical Approach.** Blackwell, Oxford.

Ferguson R 1977 **Linear regression in Geography. CATMOG** 15, Norwich.

Fingleton B 1981 'Log-linear modelling of geographical contingency tables', **Environment and Planning A** 13, 1539-1551.

Fingleton B 1984 **Models of category counts.** Cambridge University Press.

Hammond R and McCullagh PS 1978 **Quantitative Techniques in Geography: An introduction.** Oxford University Press, Oxford.

Johnston RJ 1976 **Classification in Geography. CATMOG** 6, Norwich.

Johnston RJ and Semple RK 1983 **Classification using information statistics. CATMOG** 37, Norwich.

Kirby RD 1985 **Choice in Field Surveying. CATMOG** 41, Norwich.

O'Brien LG 1989 **Generalised linear modelling in Geography.** Routledge.

Openshaw S I983 **The modifiable areal unit problem. CATMOG** 38, Norwich.

Pickles A I985 **Introduction to likelihood analysis. CATMOG** 42, Norwich.

Silk J 1981 **The analysis of variance. CATMOG** 30, Norwich.

Wrigley N 1976 **Introduction to the use of logit models in geography. CATMOG** 10, Norwich.

Wrigley N 1985 **Categorical data analysis for geographers and environmental scientists.** Longman.

10.8 Other references.

Ehrenberg ASC 1982 **A Primer in Data Reduction.** Wiley.

Hanushek EA and Jackson JE 1977 **Statistical methods for social scientists.** Academic Press, New York.

Lindley DV 1965 **Introduction to Probability and Statistics from a Bayesian Viewpoint.** Cambridge.

Moser CA and Kalton G I97I **Survey methods in Social Investigation.** Gower.

Pindyck RS and Rubinfeld DL 1976 **Econometric models and economic forecasts.** McGraw-Hill Kogakusha.

Smith SJ I984 'Crime and the structure of social relations', **Transactions of the Institute of British Geographers** 9, 427-442.

Stevens SS 1946 'On the theory of scales of measurement' **Science** 103, 677-680.

Listing of Catmogs in print

30:	Silk, The analysis of variance..	3.50
31:	Thomas, Information statistics m geography.	3.00
32:	Kellerman, Centrographic measures m geography.	3.00
33:	Haynes, An introduction to dimensional analysis for geographers.	3.00
34:	Beaumont & Gatrell, An introduction to 0-analysis.	3.50
35:	The agricultural census - United Kingdom and United States.	3.00
36:	Aplin, Order-neighbour analysis.	3.00
37:	Johnston & Semple, Classification using information statistics.	3.00
38:	Openshaw, The modifiable areal unit problem.	3.00
39:	Dixon & Leach, Survey research in underdeveloped countries.	5.00
40:	Clark, Innovation diffusion: contemporary geographical approaches.	3.00
41:	Kirby, Choice in field surveying.	3.00
42:	Pickles, An introduction to likelihood analysis.	4.00
43:	Dewdney, the UK census of population 1981.	5.00
44:	Pickles, Geography and humanism.	3.00
45:	Boots, Voronoi (Thiessen) polygons.	3.50
46:	Fotheringham & Knudsen, Goodness-of-fit statistics.	3.50
47:	Goodchild, Spatial autocorrelation.	3.50
48:	Tinkler, Introductory matrix algebra.	4.00
49:	Sibley, Spatial applications of exploratory data analysis.	3.00
50:	Coshall, The application of nonparametric statistical tests in geography	7.50
51:	O'Brien, The statistical analysis of contigency table designs	3.50
52:	Bracken, Higgs, Martin and Webster, A classification of geographical information systems literature and applications	5.00

Further titles in preparation

Order (including standing orders) from:

Environmental Publications,
University of East Anglia,
Norwich NR4 7TJ.

Prices include postage

Designed by Robert Kay and Rosie Cullington.