INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 9020812

Latent Class Analysis with partially cross-classified data

Stewart, William N., Ph.D.

City University of New York, 1990

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

A

LATENT CLASS ANALYSIS WITH PARTIALLY CROSS-CLASSIFIED DATA ..

by

WILLIAM N. STEWART

A dissertation submitted to the Graduate Faculty in Educational Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1990

This manuscript has been read and accepted for the Graduate Faculty in Educational Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

1/8/90 — Date

David Rindskopf — Chair of Examining Committee

1/8/90 — Date

Executive Officer

Dr. David Rindskopf

Dr. Alan Gross

Dr. Roger Millsap

Supervisory Committee

The City University of New York

Abstract

LATENT CLASS ANALYSIS WITH PARTIALLY CROSS-CLASSFIED DATA

by

William N. Stewart

Advisor: Professor David Rindskopf

Latent Class Analysis (LCA) separates observations in a fully-observed cross-classified table into two or more latent classes bases upon the assumption of conditional independence of the observed variables within each latent class. Frequently, a substantial number of observations are missing categorization with respect to one or more of the observed variables. We consider an alternate parameterization of the usual maximum-likelihood method for LCA in which multiple partially-crossclassified tables, that have some observed variables either Missing Completely at Random (MCAR) or Missing at Random (MAR), can be used with a completely cross-classified table or by themselves to derive the latent classes. Two specific populations are modelled and the substantial gain in using partially cross-classified tables in LCA is demonstrated.

ACKNOWLEDGEMENTS

I thank Professor David Rindskopf for the theme of this research, for aiding and guiding me in development of the analytical techniques, for redirecting me when I went astray, and for giving me a needed shove when I stalled.

I also thank my wife, Nancy, and my children - Perryn, Garret and Marla - for their patience in not having a husband/father on weekends and vacations for the better part of six years during which I completed my studies at CUNY.

Lastly, I thank the persons who have served as my intellectual models - my grandfather, Angus MacDonnell, who was a highly self-educated journalist and naval architect; my uncle, Dr. John MacDonnell, who was a school superintendent and college administrator; and my dad, Ralph, who was a long-time school principal.

TABLE OF CONTENTS

	page
Introduction	1
1. Types of Missing Data	3
2. Latent Class Analysis	6
3. Methods of Computation and Alternative Parameterization	12
Litterature Relating to Estimating of Categorical Models With Missing Data	17
Methods	27
Results	34
1. Populations/samples	34
2. Job setup	35
3. Simulation results	37
4. Generalized Linear Models	40
5. MAR tables	46
Discussion	50
Appendix I	59
Appendix II	86
Appendix III	98
Bibliography	110

LIST OF TABLES

		page
Table I	Example of incompletely cross-classified data	4
Table II	Example of partially cross-classified data	5
Table III	Probabilities of income response for three types of missing data	6
Table IV	Piagetian stage vs. height: age not accounted	7
Table V	Piagetian stage vs. height: age accounted	8
Table VI	Population parameters for Latent Class models	32
Table VII	Simulation results	38
Table VIII	Design Matrix for a 4 variable 2 latent class model, any observed variable MAR with respect to variable D	43
Table IX	"C" matrix for deriving observed cell counts for a MAR Latent Class model	44
Table X	Conditional probabilities computed via MLLSA for population #2 with variable A MAR with respect to variable D	47

LIST OF TABLES (continued)

page

Table XI Conditional probabilities computed via MLLSA for population #2 with variable A MAR with respect to variable D 49

Table XII Conditional probability cinfiguration for a series of two variable supplementary tables 51

Table XIII Discrimination of observed variables 53

INTRODUCTION

Data analysts are frequently faced with datasets which contain observations with incomplete or missing data. For instance, in assessing the relationship between college grade point average and occupational success a continuous covariate such as "years since graduation" or a categorical covariate such as sex of a subject may have been unavailable, overlooked, or obviously misrecorded. This problem oftentimes becomes troublesome as the study encompasses more variables and as the investigator(s) is distanced from the subjects either in space and time. A narrowly-defined well-controlled study on the influence of two reward conditions on math grades in four third grade classes with covariates including sex (male/female) and father's occupation (blue collar, professional, managerial) is less likely to be plagued by missing data than, for instance, a national study on the relationship between teachers's attitude and student's achievement conducted by a mailed survey wherein a teacher may refuse to answer certain questions or carelessly fail to respond to the last page of items.

Traditionally the method of dealing with an observation with a missing component has been to "throw away" the entire observation in the analyses. This method is reinforced by the major statistical packages such as SAS (Statistical Analysis System), BMDP (Biomedical Data Programs), and SPSS (Statistical Package For The Social Sciences). For example, SAS will delete all observations with a missing value on at least one of the dependent measures in the General Linear Models Procedure, oftentimes reducing greatly the effective size of the dataset. If twenty dependent variables are being analyzed and each of 1000 subjects has data on 19 out of 20 variables, then the effective sample size

for all variables is zero! A recently added procedure to BMDP (BMDPAM, 1984) does estimate missing values in a dataset using some elementary methods; these can then be utilized in other procedures.

The purpose of this study is to investigate the use of more appropriate methods of analyzing data with missing values in the case of Latent Class analysis. Two general results that apply to both continuous and categorical data when observations are deleted due to a missing component are:

(1) *Bias* in parameter estimates (that is, the expected value of the sample estimate may not be the population parameter) due to the non-random deletion of data. For example, a high school student raised in a non-permissive family environment may tend to fail to respond to a question about sexual frequency on an AIDS attitude survey; assuming that the frequency of sex would tend to be low, this omission would bias an estimate of overall frequency of sex in the group surveyed.

(2) *Loss of efficiency* in parameter estimation (that is, the standard errors of the estimates are inflated) which means that a larger sample size is required to demonstrate statistically the existence or non-existence of relationships among variables in the data.

A third effect of missing data which is relevant to Latent Class Analysis is the effect on identification status of the model. This relates to the question of whether a single set of parameter estimates will fit the data to the model.

Various techniques have been developed over the years to incorporate observations with missing data involving continuous data, originating with a technique published

in 1930 by R. A. Fisher. Over the last two decades techniques have been developed to account for missing data in categorical analyses. Both of these cases are presented in Little and Rubin (1987). However, to date, there have appeared no papers in the literature indicating how to proceed with Latent Class Analysis in the presence of missing data, although D. Rindskopf (1989) suggests that composite link functions (Thompson and Baker, 1981) in a generalized linear models framework (McCullagh and Nelder, 1983) can be utilized to do this.

1. Types of Missing Data

Missing data can be categorized according to three general patterns and to three general underlying mechanisms that are responsible for the missing data. The three patterns are incomplete, partially missing, and partially cross-classified data. The three mechanisms are missing completely at random (MCAR), missing at random (MAR), and nonignorable nonresponse.

Ploog (1967) presented an example of an *incomplete* cross-classification, (see table I) which gives a distribution of genital display among members of a colony of six squirrel monkeys (labeled as R, S, etc.). For each display there is an active and a passive participant, but a monkey never displays towards himself. Therefore by the nature of the problem the cells along the main diagonal are not observed. This "incompleteness" is mandated by the nature of the problem. Bishop, Feinberg and Holland (1975) who term such cells as "structural zeroes" give methods for proper analysis of this type of data in fully observed tables. It is easy to imagine other possible patterns of incompleteness wherein certain cross-classifications are impossible to observe.

Table I

Example of incompletely cross-classified data

active participant	Passive Participant R	S	T	U	V	W
R	-	1	5	8	9	0
S	29	-	14	46	4	0
T	0	0	-	0	0	0
U	2	3	1	-	38	2
V	0	0	0	0	-	1
W	9	25	4	6	13	-

Data having different levels of specification for a categorical variable are said to be *partially missing* (*partially specified* might be a less confusing term!). For example, in a survey some subjects may be asked a long form of the questionaire which for the question on marital status includes categories of (a) married and living together (b) married but separated (c) divorced (d) widowed (e) never married; a short form of the questionaire will include responses of only (a) married or (b) not married.

Partially cross-classified data, refers to the situation where some subjects have no response for one or more of the cross-classificatory variables. Little and Rubin (1987) provided a fabricated but realistic example of a simple problem involving this situation data (table II). Here there is a table of completely cross-classified data and a second table in which units are only classified on the variable Y1. The second table represents additional information with respect to the marginal distribution of Y1.

Table II

Example of partially cross-classified data

	complete units			incomplete units	
		Y2			
	1	2	3		
Y1 1	20	30	40	1	100
Y1 2	50	60	20	2	90

data are said to be *missing completely at random* (MCAR; Little and Rubin,1987) when the probability of response is independent of the classificatory variables. For example, for the two variables age and income, if the probability that income is recorded is the same for all individuals regardless of age or income, then the data are MCAR. This is illustrated in table IIIa where the probabilities of response are constant across all cells.

The data are *missing at random* (MAR) if the probability of response is dependent upon observed values of a variable. If the probability that income is recorded varies according to the age of the respondent, and age is observed for everyone, but not to the income level within each age category, then the data are MAR. Table IIIb illustrates MAR. The assumption of MAR is most useful in that likelihood-based methods can often be applied in data analysis, e.g., the likelihood can often be factorized (Little and Rubin, 1987).

Table III

Probabilities of income response for three types of missing data

	(a) (MCAR) age			(b) (MAR) age			(c) (NINR) age		
	20-40	40-60	>60	20-40	40-60	>60	20-40	40-60	>60
income (x1000)									
<20	.8	.8	.8	.9	.8	.7	.9	.8	.7
20-50	.8	.8	.8	.9	.8	.7	.6	.5	.4
>50	.8	.8	.8	.9	.8	.7	.3	.2	.1

Nonignorable nonresponse (NINR) occurs when neither MCAR nor MAR holds. For the age-income example, if the probability of response for the variable income depended upon the income level within each age category, then the nonresponse would be nonignorable. For example, within an age category if those with high income were less likely to respond (i.e., less likely to provide information on income) then nonresponse is not ignorable. Table IIIc illustrates NINR.

2. Latent Class Analysis

Modern models for latent class were developed by Goodman (1974) based upon earlier development by Larzarsfeld (1968). They have proved useful in a variety of educational research problems including identification of mastery and non-mastery

classes in criterion-referenced testing (Rindskopf, 1984); identification of heirarchical learning skills(Rindskopf, 1984); measuring agreement when the categorical response are subject to error (Bergan, 1982); and determining response consistency accross different groups of individuals, e.g., in assessing test bias a test developer may wish to find out whether mastery classifications for item domains are reliable in different ethnic groups (Bergan, 1982).

Suppose Table IV (modified from Rindskopf, 1983) were presented for analysis . One might ask whether there a relationship between height and Piagetian stage. The ratio in the crossproducts of the diagonals is not 1 supporing this hypothesis; there is a relationship of height and stage. However a second question of interest is whether there is a third variable which could account for the the apparent observed relationship. In other words, given the value of the third variable does the observed relationship disappear?

Table IV

Piagetian stage vs. height: age not accounted

		non-conservers	conservers
height	tall	28	52
	short	52	28

Table V shows that two subtables can be constructed from Table IV in which height and Piagetian stage are *independent*. The relative proportions of non-conservers to conservers for tall and short students is the same within a subtable (4:1 in the subtable [a] vs 1:4 in subtable [b]). If the ages of the children were checked it might be seen that all the children in subtable (a) were 7 years old and all the children in subtable (b) were 9 years old. The researcher would then be in a position to state that age was the likely explanatory cause for conservation status, especially if it were supported by theory.

Table V

Piagetian stage vs. height: age not accounted

	subtable (a)		subtable (b)	
	non-conserver	conserver	non-conserver	conserver
height				
tall	16	4	12	48
short	48	12	4	16

The purpose of Latent Class Analysis (LCA) is to find unobserved (i.e., *latent)* variables that, like age in the example above, would explain the relationships among observed categorical variables. LCA is the categorical analogue of factor analysis. Factor analysis assumes that variables, both manifest and latent, are continuous, whereas LCA assumes that the variables are categorical and multinomially distributed. In some cases where dichotomized data are presented, and it is known that the dichotomized variables have been created from originally continuous variables, Factor

Analysis can be performed on the tetrachoric correlation matrix (Bock and Lieberman, 1970; Muthén, 1979). The crucial assumption of those techniques is that of *conditional independence,* which states that given the values of the unobserved (latent) variables the observed (manifest) variables are independent of one another. The classical presentation of the latent class model (Goodman, 1974, 1978) will be given below; an alternative representation in terms of Generalized Linear Models (McCullough and Nelder, 1983) will be given later.

Symbolically, let $\underline{A}_i$ (i=1 to I), $\underline{B}_j$ (j=1 to J), $\underline{C}_k$ (k=1 to K), etc., identify categories of P observed variables $\underline{A}$, $\underline{B}$, $\underline{C}$, etc., and let $\underline{X}_t$ identify T classes of a latent variable called $\underline{X}$. Furthermore let:

$\pi^{ABC}_{ijk...}$ = the probability that an individual will be at level (ijk...) with respect to the observed variables $\underline{A}$, $\underline{B}$, $\underline{C}$, etc.

π^{X}_{t} = the probability that an individual is in class t of the latent variable

$\pi^{\overline{A}X}_{it}$, $\pi^{\overline{B}X}_{jt}$, etc. = the conditional probability that an individual will be at level i of variable $\underline{A}$ given that the individual is in latent class t, etc.

Then:

$$\pi^{ABC...X}_{ijk...t} = \pi^{X}_{t}\,\pi^{\bar{A}X}_{it}\,\pi^{\bar{B}X}_{jt}\,\pi^{\bar{C}X}_{kt}\ldots \qquad (1)$$

$$\pi^{ABC}_{ijk...} = \sum_{t=1}^{T}\pi^{ABC...X}_{ijk...t} \qquad (2)$$

and substituting (1) into (2):

$$\pi^{ABC}_{ijk...} = \sum_{t=1}^{T}\pi^{X}_{t}\,\pi^{\bar{A}X}_{it}\,\pi^{\bar{B}X}_{jt}\,\pi^{\bar{C}X}_{kt}\ldots \qquad (3)$$

By definition:

$$\sum_{1}^{T}\pi^{X}_{t}=1\,,\quad \sum_{1}^{I}\pi^{\bar{A}X}_{it}=1\,,\quad \sum_{1}^{J}\pi^{\bar{B}X}_{jt}=1\,,\ \text{etc.} \qquad (4)$$

Equation (2) states that individuals can be classified into T mutually exclusive and exhaustive latent classes and equation (1) states that within the t^{th} latent class the manifest variables are mutually independent. The objective of LCA is to obtain parameter estimates of the π^{X}_{t}'s and $\pi^{\bar{A}X}_{it}$ etc. so that equation (2) is properly fit.

Goodman (1974) provides an iterative procedure for maximum likelihood estimation of the parameters, which is an example of an EM (*Expectation-Maximization*) algorithm (see below) for deriving parameter estimates.

Everitt and Hand (1981) and Everitt (1984, 1987) provide an alternate presentation of the latent class model in terms of a finite multinomial mixture density. This is fully equivalent to Goodman's presentation.

A third presentation of the model in terms of Generalized Linear Models will be given below; the GLM approach is computationally different and can handle a wider array of models .

A major concern in performing LCA is whether the model is *identified.* Identification relates to the concept of having a unique set of parameter estimates that are consistent with the population model. If different sets of parameter estimates imply the same $\hat{\pi}^{ABCD}_{ijkl}$ then the model is said to be unidentified and, by implication, parameter estimates are not directly interpretable. A simple analogy of an unidentified model in a non-LCA situation is the case where a set of linear simultaneous equations is to be solved but there are more unknowns than equations; in this case the solution set is not unique - there are many sets of solutions to the equations.

In LCA the minimum necessary condition that an unrestricted model is identified is that:

$$I*J*K*... > (I + J + K +... - (p-1))*T \qquad (5)$$

is true (p=the number of manifest variables). This equation states the there should be more observed proportions than there are parameters estimated by the model. A second and definitive condition for indentifiability that is more computationally difficult is

that the Jacobian matrix (i.e., the matrix of derivatives of the $\pi_{ijk...}^{ABC}$ with respect to the parameters) of equations (3) have a rank equal to the number of columns in the matrix (see Goodman, 1974). This condition can be tested numerically following parameter estimation.

A further refinement of the model is the ability to put restrictions on the parameters. If it is known or suspected in advance that two or more parameters are equal then these equalities can be forced during the analysis, resulting in the need for fewer parameter estimates. In addition, parameters can be fixed at specific values, thus eliminating altogether the need to estimate these parameters. Sometimes, with fewer parameters that need to be estimated, a previously unidentified model will become identified.

Pearson or likelihood ratio goodness-of-fit statistics are used to test adequacy of the model. Both statistics approximate a chi-square variable when the model is true. In both methods the discrepancy between actual frequencies in each observed cell of the table and frequencies predicted by the model is tested. The degrees of freedom are calculated as the number of independent probabilities in the observed table minus the number of independent parameters estimated.

3. Methods of Computation and Alternative Parameterization

There are two primary methods of performing LCA. Both are iterative in the sense that closed-form estimates of the parameters cannot be attained. Instead, initial estimates of the parameters must be supplied to each procedure and then the procedure iterates until it converges to a final solution or it diverges to no solution. Oftentimes the initial estimates are quite different from the final estimates, but there are "good"

and "bad" starting values. The degree of success in obtaining a final converged solution often depends upon one's ability to find good starting parameter estimates.

The "traditional" method (Goodman, 1974) was programmed and distributed (Clogg, 1977) under the name MLLSA (*Maximum Likelihood Latent Structure Analysis*). The basic computational algorithm uses starting parameter estimates of the latent class probabilities (the π_t^X) and of the marginal conditional probabilities (the $\pi_{it}^{\bar{A}X}$, $\pi_{jt}^{\bar{B}X}$, etc). First equations (1) and (2) are calculated with the vector of initial parameter estimates. Then the latent class probabilities and marginal conditional probabilities are re-estimated by using equations (6), (7) and (8) below. Note that "$p_{ijk...}$" is the observed proportion in cell(ijk..).

$$\hat{\pi}_{ijk...t}^{ABC...\bar{X}} = \hat{\pi}_{ijk...t}^{ABC...X} \; / \; \hat{\pi}_{ijk...}^{ABC...} \qquad (6)$$

$$\hat{\pi}_t^X = \sum_{ijk...} p_{ijk...} \, \hat{\pi}_{ijk...t}^{ABC..\bar{X}} \qquad (7)$$

$$\hat{\pi}_{it}^{\bar{A}X} = \left(\sum_{jk..} p_{ijk...} \, \hat{\pi}_{ijk...t}^{ABC...X} \right) / \; \hat{\pi}_t^X \qquad (8)$$

Equations similar to (8) can be derived for B, C, etc. Equation (6) states that the probability of being in class (t) given that an individual is in the i^{th} category of

A and j^{th} category of B and k^{th} of C etc. is equal to the probability of being in the i^{th} category of A and j^{th} category of B ... and t^{th} class of X divided by the probability of being in the i^{th} category of A and j^{th} category of B and k^{th} category of C, etc. With the new parameter estimates obtained from equations (7) and (8) a second iteration is started and so on. The procedure stops when either a prespecified number of iterations is reached (no convergence) or when the change in the chi-square from the observed cell frequencies and those predicted by the model is small enough. In addition the program will do an identification check by calculating the rank of the Jacobian matrix of equations (3) (Goodman, 1974).

The MLLSA algorithm is an EM type algorithm (Dempster, Laird & Rubin, 1977). Each iteration consists of two basic steps: an *expectation* step in which the the "missing" data, i.e., the $\pi_{ijk...t}^{ABC...X}$ (the unseen cell proportions in each of the latent classes including latent as well as observed variables) are estimated using the current set of parameter estimates (equation [1]), and a *maximization* step in which the estimated complete data are used, as if they were observed data, to compute revised parameter estimates (equations [7] and [8]).

Another approach to fitting Latent Class models is by using a *Generalized Linear Models* (GLM) approach (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1983). This approach is a very generalized approach that can fit linear models to data from any distribution in the exponential family, including the multinomial models of LCA. There are three basic components to a GLM:

1. A random component which specifies the distribution of the dependent variable $\underline{Y}$. The expected value of $\underline{Y} = \mu$.

2. A systematic component consisting of covariates x_1 , x_2 ... , x_p which produce a linear predictor η given by:

$$\eta = \sum_{1}^{p} \beta_j x_j \quad = \quad X\beta$$

where the βs are parameters whose values are usually unknown and have to be estimated from the data, and the x_j form what is generally termed the *design* or *model matrix.*

3. The *link* function between the random and systematic components:

$$g(\mu) = \eta$$

For contingency table models the link function is often the logarithmic function, $\log(\mu) = \eta$.

Some types of data, however, do not fit directly into a GLM model. For example, if the observed frequencies in observed categories were actually obtained by combining multiple underlying (unobserved) categories, then the model has to account for the multiple sources of the observed frequencies. One example is that of the O, A, B, and AB blood groups (phenotypes) which arise by combining various cells of the underlying nine genotypes obtained from the A, B, and O genes from the mother and the father. Another example is LCA where the observed table is hypothesized to have arisen by the combining of two or more unobserved subtables from the latent classes.

The extension of GLM to account for the more complex cases mentioned above was proposed by Thompson and Baker (1981). They included the possibility of having multiple link functions in the model which they termed *composite link functions.* Specifically, for linear functions, they parameterized the model as follows:

$$\mu = \underline{C}\gamma$$

where: $\gamma = h(\eta)$

Here γ is a system of link functions and $\underline{C}$ is a matrix that indicates which of the (unobserved) cells produced by the link functions are to be combined to obtain the observed set of cells.

If the link function is logarithmic then :

$$\mu = \underline{C}e^{(X\beta)}$$

It turns out that the composite link function approach can not only be applied to estimating the latent classes in LCA but can also be used in a variety of missing data situations. More will be said later about this.

LITERATURE RELATING TO ESTIMATING OF CATEGORICAL MODELS WITH MISSING DATA

An early paper pertaining to analysis of incomplete discrete data was that of Hartley (1958). He gives an iterative method for obtaining maximum likelihood estimates from truncated or censored data. For a variable that follows a Poisson distribution the procedure is as follows:

1) estimate initial values for the missing frequencies, e.g., if the frequencies for zero and one counts are missing then supply an initial estimate of them;

2) using both the actual and estimated data, estimate the mean (the parameter, θ) from the maximum likelihood equation appropriate to the complete sample (this is the maximization step in the EM algorithm);

3) using the parameter estimate, compute "improved" estimates of the of the missing frequencies by proportional allocation based on Poisson distribution, i.e., $n_j = nf(j,\hat{\theta})/((1-f(j,\hat{\theta}))$, where n_j is one of the missing counts, n is the total number of non-missing data values, and $f(j,\hat{\theta})$ is the estimated Poisson probability for the j^{th} missing count (this is the expectation step in the EM algorithm);

4) iterate (2) and (3) until there is little change in the estimated mean from iteration to iteration.

Hartley found the convergence to be extremely rapid, generally occuring within five cycles, although poor starting values increased the number necessary.

Blumenthal (1968) presented methods when data were partially categorized under multinomial sampling. The particular sampling situation Blumenthal is concerned with is that of 2 or more main categories with 2 or more subcategories within each main category; sometimes it is difficult to distinguish among the subcategories and therefore the subcategories are not distinct for some observations. The author assumes that a sample belonging in the j^{th} subcategory of the i^{th} main category has probability α_{ij} of being only partially classified. The α_{ij} are incorporated in equations of the multinomial model and probability estimates are derived directly without iteration.

The categorical linear model approach of Grizzle, Starmer and Koch (1969) was extended to cross-classifications with partially crossclassified data by Koch, Imrey and Reinfurt (1972). They give an example where the majority of the data are from a 3-factor design with two levels at each level (success=1 or failure=2) with supplementary marginal tables for each 2-factor combination and for each of the single factors. They solve the following weighted least squares model for estimates of π_1, π_2 , and π_3 which are success rates of each of the three factors (treatments) :

$$A * p_G = F = X \hat{\pi}$$

$$\hat{\pi} = (X' V_F^{-1} X)^{-1} X' V_F^{-1} F$$

where: A is the following 12 x 26 matrix:

1 1 1 1 0 0 0 0
1 1 0 0 1 1 0 0 (for complete table)
1 0 1 0 1 0 1 0

1 1 0 0 0's elsewhere
1 0 1 0

1 1 0 0 (for supplementary 2-way
1 0 1 0 tables)

1 1 1 0
1 0 1 0

1 0

1 0 (for supplementary 1-way
1 0 tables)

p_G is 26 x 1 vector containing the probabilities of response for each cell in the complete table and for each cell in each of the supplementary tables, i.e. ("." denotes collapsing),

$$p_G = (\hat{\pi}_{000}, \hat{\pi}_{100}, \hat{\pi}_{010}, \hat{\pi}_{110}, \hat{\pi}_{001}, \hat{\pi}_{101}, \hat{\pi}_{011}, \hat{\pi}_{111}, \hat{\pi}_{00.}, \hat{\pi}_{10.}, \hat{\pi}_{01.}, \hat{\pi}_{11.}, \hat{\pi}_{0.0}, \cdots \hat{\pi}_{..1})$$

$\hat{\pi}$ is a 3 x 1 vector of estimates of the factor success probabilities in the combined complete table,

X is a 3 x 12 design matrix which contains an identity matrix of rank 3 and additional rows for each of the parameters that

appears in each of the supplementary tables, i.e.,

$$X = \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}$$

Note, for example, rows 5 and 6 are rows representing the supplemental A x B table and row 12 is the supplemental C table.

V_F^{-1} is the inverse of a function of covariance matrix for the 8 cell proportion estimates derived from the complete table plus the 18 additional marginal proportion estimates derived from the marginal samples; $V_F = A[V(p_G)]A'$ where A is defined above $V(p_G)$ is the covariance matrix of p_G.

Koch et. al go on to show how a χ^2 goodness-of-fit test can be constructed and

how various hypotheses, such as $\pi_1 = \pi_2 = \pi_3$ can be tested on the full estimated table.

Hocking and Oxspring (1974) gave a somewhat similiar account using a maximum likelihood approach. As in Koch et. al. the authors use matrices containing indicators (0's and 1's) of which of the available cell proportions are to be combined to form the estimates of the population cross-classified proportions. They set up the basic likelihood equation as follows:

$$W \pi = \sum_{t=1}^{T} D_t' W_t P_t$$

where: π_n = population proportions

P_1 = cell proportions in complete table

$P_2, \ldots, P_t$ = proportions in supplemental tables

$D_1 = I_{n-1}$ = identity matrix with dimension equal to 1 less than the number of population proportions

$D_2, \ldots, D_t$ = matrices with n-1 columns and s-1 rows indicating which proportions in the complete table were combined; s = dimension of the supplementary table

W = variance-covariance of the π_n

W_t = variance-covariance of the cell probabilities in table (t)

Note that the W and W_t matrices act as weights; without them every cell would be weighted equally, but the variance of a proportion is related to the magnitude of the proportion.

Hocking and Oxspring describe how to test the hypothesis $G\pi = 0$, where G is the design or hypothesis matrix, by iteration. The equation to be solved is:

$$\hat{\pi} = (I - W^{-1}G' (GW^{-1}G)^{-1} G) W^{-1} \sum_{t=1}^{T} D_t' W_t P_t$$

Both of the above approaches produce estimators that are asymptotically normal, i.e., with large same sizes the estimators are approximately normally distributed and therefore hypothesis testing can use normal theory. A difference between the two is that the Hocking-Oxspring method derives estimators from the maximum likelihood method and assumes a product-Poisson or product-multinomial distribution whereas the Koch et. al. method makes fewer assumptions about the underlying distribution. The advantage of using maximum likelihood methods over weighted least squares is that a smaller sample size is required to support the asymptotic properties of the estimated parameters (Koch and Imrey, 1985 [anal. of cat. data]) .

Chen and Feinberg (1976) obtained multinomial maximum likelihood estimates for the case in which some cells in a contingency table were indistinguishable. The estimates of row marginal counts are obtained as the sum of the observed counts in a row plus a proportional allocation of the relevant totally mixed-up counts. Identifiability problems arise when (a) two or more columns or rows are completely mixed up in rectangular block forms, and (b) the decrease in degrees of freedom due mixed-up frequencies is greater than (I-1)(J-1) where there are I rows and J colums

in a two-way table.

Dempster, Laird and Rubin (1977) applied the EM algorithm to compute maximum likelihood estimates from incomplete data. As mentioned previously, each iteration consists of two basic steps: an *maximization* step in which conditional probabilities are calculated from current estimates of the complete cell probabilities rather than just from the fully classified data. The *expectation* step then calculates new cell probabilities from the estimated completed data. The basic algorithm for contingency tables (the algorithm can be applied to other types of data, i.e., continuous) proceeds as follows:

Step 1: Calculate conditional probabilities from the estimated complete table obtained from the previous iteration, or if the first iteration then use the complete table. E.g., if the complete table is 2 x 2 and there are supplemental marginal tables for both margins from the completed table, estimate $p(A_1 | B=1)$, i,e., the probability that variable A is at level 1 given that variable B is at level 1.

Step 2: Multiply the supplemental frequency at level 1 of B in the marginal table that is collapsed over A and add this to the estimated full table frequency from the previous iteration.

Step 3: Do (1) and (2) for each of the supplementary tables.

Step 4: Re-estimate the complete table cell probabilities and repeat steps (1) - (3) until the change in estimated complete cell

probabilities changes little with additional iteration.

The algorithm is a method relating to traditional thinking about handling missing data, namely: (1) replace missing values by estimated values, (2) estimate parameters, (3) re-estimate the missing values assuming the new parameter estimates are correct, (4) re-estimate parameters, (5) iterate until convergence.

Little and Rubin (1987) state that the algorithm converges reliably to a local maximum or saddle point of the loglikelihood $l(\theta|Y_{obs})$, i.e., the loglikelihood of the parameters given the observed data. A disadvantage of the algorithm is that the convergence rate is inversely related to the amount of missing data and can be very slow when large amounts of data are missing. An assumption of the method is that the data are MAR (missing at random; see above) and that the complete data counts have a multinomial distribution.

Fuchs (1982) applied the EM algorithm to data from the Protective Services study, which had data from a complete six-way table and three subtables where one or two of the variables was not categorized. The manner in which the data were originally analyzed was to collapse the tables over the variables that had missing data so that all the observations could be used in a four-way analysis of the data.

Fuchs estimated the complete table via the EM algorithm and then submitted the data to a complete six-way loglinear analysis where the primary relationship between disease and group membership (experimental or control) was of key interest. In the collapsed four-way table the original authors found a significant association, but re-analysis of the estimated six-way table by Fuchs found no such relationship. He blames the difference on the confounding of covariates not originally accounted for

in the model used by the original authors.

Nordheim(1984) developed a method for including "fused" data into categorical data analysis when the MAR (see above) assumption does not hold, i.e., the "fused" or missing data that is not assigned to specific categories is non-randomly missing. Alternatively, the mechanism of missing data is said to be non-ignorable. Nordheim assigns probabilities of a value being missing dependent upon the true category to which the value belongs and incorporates these probabilities in the likelihood function. He applies the method to a problem involving Turner's Syndrome where the probability that a physician records that a symptom, such as webbed neck, is present when it truly is present is higher than than the probability that the symptom is recorded as absent when it is truly absent. In other words, the category of "missing" (neither present nor absent, the physician did not record a value for a patient) is more likely to contain a higher proportion of true "absents" than the ratio of true "absents" to true "presents" would indicate.

Woolson and Clarke (1984) described the use of weighted least squares for incomplete longitudinal categorical data. Rather than augmenting the X (design) matrix with additional blocks for each supplemental table as Koch, Imrey and Reinfurt (1972) suggested they augment the A (function of cell probabilities) matrix by addition of responses with a missing value (e.g., M-O-N for missing at first time point, obese at second time point, and not obese at third time point) and by addition of columns, one for each year, which have a '1' if a value was recorded for that particular year or a '0' if the value was missing. The authors note the distinction between their method and that of Koch, Imrey and Reinfurt but present no comparison of the two approaches. The authors show that the parameter estimates using all the data are not much different than the parameter estimates using only the complete data; however, the variances of the estimates are considerably smaller (about half of those of the complete data).

Rindskopf (1989) indicates how to use generalized linear models with composite links to fill in contingency tables with supplementary margins, fit loglinear models to contingency tables with missing data, fit latent class models without and with missing data on the observed variables, and to fit models with various assumptions about the missing data process. He provides examples of filling in contingency tables with supplemental margins and performing loglinear modeling with missing data. The author specifies the form of the X (design) matrix and the C (link) matrix in each case needed to perform generalized linear model analysis. In each case the Y (response) vector is augmented by the supplemental data, the X matrix is augmented by duplicating itself for each subtable and additional columns are added to the X matrix (assuming MAR) as indicators of to which table (complete table, supplementary table [A], supplementary table [B], etc.) the responses belong. The C matrices consist of (1) an identity submatrix with rank equal to the number of cells in the complete table and (2) submatrices indicating which cells of the complete table are combined to form the supplementary tables; there is a submatrix for each supplementary table.

METHODS

Three different methods of including tables with partially missing data in Latent Class Analysis were tested on datasets with four different levels of missing data, ranging from the complete dataset only to analyses including only partially missing, in each of two different populations. To perform the data analyses two programs were written in SAS IML; one program performed Generalized Linear Models analysis and the other did Maximum Likelihood Latent Structure Analysis with fully-crossclassified and partially-crossclassified data.

The underlying model assumed for all of the analyses is a multinomial mixture model of the form of equation (2). In the most basic terms this model states that the observed proportion in a particular cell of the observed crossclassified table is a weighted sum of the same cells of each of the t latent classes. The weights are the probabilities of belonging to the latent classes and within each latent class the probabilities of being in a particular cell is simply the product of probability of marginal membership for each of the crossclassificatory variables, i.e., the observed variables are independent within each latent class.

For the observed supplementary tables it was assumed that the data were MCAR or MAR. For each cell of the supplementary table(s) this means that the original allocation of these frequencies was proportional to the the relative frequencies of the non-missing variables in the complete table, e.g., if we reallocate the collapsed frequencies then the frequencies in the original cells, say for the three-way table Piagetian stage example (page 7) where there is an observed supplementary one-way table in variable i that is

collapsed with respect to variables j and t, would be:

$$n_{ijt} = n_i * \pi_{i \mid jt}$$

The MAR assumption allows that "missingness" is conditional on which i^{th} category an individual belongs in proportion to the jt^{th} class over which the data has been collapsed. In practical terms this means that the marginal frequencies with respect to the i's in the complete table may be relatively high whereas the frequencies of the i's in the supplementary table may be low or even zero. These ideas are used as indicated below to allocate frequencies from supplemental tables in Latent Class Analysis.

For the piagetian stage table this would indicate that although we had the conserver status of each individual (variable j) the height of some individuals (variable i) in subtables (a) and (b) (the subtables are variable t) was missing as follows:

subtable (a)		subtable (b)	
non-conserver	conserver	non-conserver	conserver
10	8	4	4

These four supplementary frequencies would be allocated to the complete table with respect to the four tall/short proportions in the subtables (a) and (b), e.g., (4/16) x 8 =2 supplementary subtable (a)/conserver subjects would be allocated to the subtable (a) /conserver/tall category and (12/16) x 8 = 6 subtable (a)/conserver subjects would

be allocated to the subtable(a)/conserver/short category. In practice, if the subtables represent a latent variable then all we would observe for a supplementary table is:

non-conserver	conserver
14	12

and a procedure must be developed for allocating these frequencies in the process of determining the latent classes.

The MAR assumption for LCA is equivalent to saying that cells in the supplemental table were derived by collapsing over cells in the complete i*j*... table and therefore frequencies in the supplemental table should be allocated to the complete table based upon the relative proportions of the cells in the the complete table that have the same subscripts as the supplementary cells. For example, for a two latent class (X_t) model where there are four dichotomous observed variables A_i , B_j , C_k, D_l the frequencies in a BCD supplementary table would be allocated as:

$$\text{nsupp}_{i\,|tjkl} = \left(\text{ncomp}_{i\,|tjkl} \,/\, \sum_{i} \text{ncomp}_{i\,|\,tjkl} \right) * \text{nsupp}_{tjkl}$$

where $\text{nsupp}_{i\,|\,tjkl}$ is the supplemental frequency to be added to cell tijkl of the observed table $\text{ncomp}_{i\,|\,tjkl}$ is the frequency of cell(tijkl) in the complete table and nsupp_{tjkl} is the frequency of cell(tjkl) in the supplemental table.

However, with LCA the "complete" table is never observed and so the procedure must be modified. For the EM algorithmn we proceed as we would normally do with LCA via Maximum Likelihood estimation (see equations 1-3 and 6-8) but during each iteration the supplementary tables would be included in the analysis by treating them as additional latent classes and applying appropriate restrictions across the classes in what Clogg has termed "simultaneous group" Latent Class Analysis. For example, for a two class four observed variable LCA with one supplementary table which is missing one of the four observed variables, say A, then the program set-up would include: (1) latent classes 1 and 2 for the completely observed table and latent classes 3 and 4 for the incompletely observed data; (2) for the variable A in the supplementary table data input would consist of marginal A frequencies collapsed into category A_1 with all other A cells having zero frequencies; (3) a fifth observed variable which indicates group membership, in this case either the complete or supplementary table; (4) restrictions on conditional probabilities across the groups except on the variable on which the MAR conditioning is occurring.

For the GLM approach SAS IML (Interactive Matrix Language) was used to create a Generalized Linear Modelling program which is able to perform GLM with composite link functions (Thompson & Baker, 1981). This program is capable of performing LCA with observed complete and supplemental tables by appropriately specifying design matrices, link functions and initial values. Unless "good" initial values are chosen, however, the procedure may fail to find a solution.

SAS IML was also used to create a Maximum Likelihood Latent Structure program by the procedure for EM estimation given above. The program is sufficiently general so that classification results for any LCA model with the "usual" restrictions (equations [4]) can be done. The published program MLLSA (Clogg, 1977) was also

used for incorporating supplementary tables. Both programs yield the same conditional and latent group probabilities but each has its own advantages, which will be discussed in the Discussion section. For both programs, as indicated above, each observed table was treated as an observed group and appropriate restrictions were applied to the conditional probabilities within and among the groups (tables). Assumptions were examined.

All three programs are documented as to how to input data, matrices, conditional probability restrictions, etc. to obtain solutions that incorporate supplemental tables. Simulations were done using the maximum likelihood approach, although solutions and interpretations for both major approaches (Maximum Likelihood and GLM) are discussed.

Populations with known parameter values were generated and then these parameter values were used as initial values with MAR supplemental tables in both MLLSA and GLM procedures. Essentially this was a check on the correctness of the programs written (GLM or EM IML) and use of the existing program (MLLSA). We consider examples with four observed variables for some subjects (variables A, B, C, and D) and for other subjects information on only observed variables B, C, and D and/or A, C, D is available. Two populations were generated from the following parameter values in Table VI.

Population #1 represents a population whose parameters are not too different from one another and therefore represent a "difficult" case for 2-class models whereas population #2 represents a population whose latent classes are well differentiated and therefore represents an "easy" case for 2-class models.

Table VI

Population parameters for Latent Class models

population #	latent class	π_t^X	$\pi_1^{\bar{A}}$	$\pi_2^{\bar{A}}$	$\pi_1^{\bar{B}}$	$\pi_2^{\bar{B}}$	$\pi_1^{\bar{C}}$	$\pi_2^{\bar{C}}$	$\pi_1^{\bar{D}}$	$\pi_2^{\bar{D}}$
1	1	.3	.4	.6	.1	.9	.6	.4	.3	.7
1	2	.7	.6	.4	.3	.7	.4	.6	.1	.9
2	1	.5	.1	.9	.2	.8	.9	.1	.8	.2
2	2	.5	.9	.1	.8	.2	.1	.9	.2	.8

Collapsing was done over variable A to form a marginal BCD table and over variable B to form a marginal ACD table. Although the simulations relate to MCAR supplementary tables the same methodology is easily applied to MAR tables with one alteration: the conditional probabilities for the conditioning variable(s) are NOT restricted across simultaneous groups. This is explained below where an example is given and interpreted.

Thirty samples of 100 observations via multinomial sampling were obtained from each of the two populations generated according to the conditional probabilities in the table above. From these thirty samples ten analytical groups were formed, each consisisting of three of the thirty samples. In each group there was a "complete" table cross-classified with respect to A, B, C, and D; a marginal supplementary table collapsed over A; and a marginal supplementary table collapsed over B. This method yields MCAR supplementary tables. A method for deriviving MAR tables through

GLM is provided in the results section.

For each of these ten analytical groups x two populations the following analyses were performed:

1) The complete table only.

2) The complete table plus B*C*D supplementary table.

3) The complete table plus B*C*D and A*C*D supplementary tables.

4) The B*C*D and A*C*D tables only, i.e., only partial data are included in the analyses.

Thus, 80 analyses were performed. To compare results among (1) through (4) above an index was constructed: because the population conditional probability parameters were known, the sum of absolute deviations (sample conditional probabilities minus population conditional probabilities) were computed and compared. The objective was to determine whether and to what degree (2) and (3) above were superior to (1) in estimating population conditional probabilities and to determine how (4) compared to (1). The usual chi-square test reported in Clogg's MLLSA program is not an appropriate measure because the estimated observed frequences are tested against the actual observed frequencies, not against the underlying population. A chi-square test of estimated observed frequencies could have been performed; however, using a measure directly involving the population conditional probabilities is more easily interpreted.

RESULTS

I. Populations/samples

MLLSA program listings for analysis of the two populations are presented as programs 1, 2, 3 and 4 in Appendix I. Programs 1 and 3 are the "complete" population data only for populations #1 and #2, respectively, whereas programs 2 and 4 are the "complete" population plus variable A and variable B MCAR populations. These are basically verifications of population and MCAR tables but they also demonstrate basic program set-up. As expected, all four analyses recapture the population conditional and latent group probabilities. The general set-up for MLLSA for single and simultaneous groups is documented in Clogg (1977). The set-up for incorporating supplementary marginal tables is documented below.

The SAS program used to obtain 30 multinomial samples for the first set of population parameters is presented in Appendix I as program 5. The output frequencies were then submitted to a second program (program 6) which processed these frequencies in ten groups of three samples with the first sample in a group of three treated as the complete A*B*C*D table, the second sample collapsed over A to form the marginal table B*C*D, and the third sample collapsed over B to form the marginal table A*C*D. This procedure was repeated for population #2.

II. Job Set-up

To use MLLSA for MCAR supplementary tables the following procedure must be used:

1) Each additional table will be entered as additional data input lines with the same number of entries as the complete table and in the same order as the complete table; for the variable that is collapsed non-missing data will be entered as the first level of the collapsed variable and zeros will be inserted for level 2 and beyond for the collapsed variable. For example, for an analysis with 4 observed variables and two levels of each variable, the data for the complete A*B*C*D, B*C*D, and A*C*D marginal tables would have the following pattern, where the order of frequencies is $A_1B_1C_1D_1$, $A_2B_1C_1D_1$, $A_1B_2C_1D_1$, etc., :

1	5	2	21	7	2	1	4	2	2	3	8	27	5	8	2	A*B*C*D
6	0	31	0	5	0	5	0	6	0	4	0	35	0	8	0	B*C*D
8	38	0	0	6	5	0	0	3	6	0	0	21	3	0	0	A*C*D

Note that there are 0's in the second line wherever the entry is an A_2 entry and there are 0's in the third line wherever there is a B_2 entry.

2) Although the model is a two latent class model, because three groups are being analyzed there is a total of 2 x 3 = 6 latent classes.

3) An additional variable is set up to define to which group a frequency belongs, e.g., for the example, a fifth variable E is set up with level 1 corresponding to the complete table in the first line of data above, level 2 corresponding to the B*C*D table on the

second line, and level 3 corresponding to the A*C*D table on the third line.

4) The same conditional probability is entered for corresponding entries in simultaneous groups for variables that are not collapsed and they are restricted to equal one another. For instance, the conditional probability for C_1 in latent classes 1, 3 and 5 is restricted to be the same. However, for variable A, the conditional probability for A_1 in class 1 would be would be restricted to be equal for the same in class 5 but for class 3 it would be set equal to 1 since for class 3 all variable A frequencies are collapsed into A_1.

5) For the variable indicating "group" , variable E in the example, a "1" is entered at level #1 for the complete group for both latent classes within that group, with 0's at other levels of the variable for that group. For the first supplementary table "1's" are entered at level #2 and "0's" elsewhere. For the second supplementary group "1's" are entered at level #3 of of the group variable and "0's" elsewhere. The pattern would be repeated for as many supplementary tables that were to be incorporated into the analysis.

The SAS program in Appendix III illustrates the principles given above. The following should be noticed:

1) The final likelihood ratio statistic is 0, indicating a perfect fit, as expected.

2) The conditional probabilities are the population conditional probabilities in each of the three groups except (a) where the variable has been collapsed where it is equal to "1" or "0", and (b) for the fifth variable, which is a group indicator variable.

3) Within each group the sum of the latent class probabilities sums to the proportion of the entire table that is accounted for by that group, i.e, if the complete table accounts for 33% of the total frequency then the sum of the latent class probabilities for the complete table will be .33.

For MAR tables the same procedure as with MCAR tables can be used with one exception: the conditional probabilities across groups for the variable that determines missingness must not be restricted. For example, if variable A is MAR with respect to variable D then the conditional probabilities for D must not be restricted.

III. Simulation Results

For the complete table, the simulation in ten 100-observation samples from population 1 indicates an average deviation from population conditional probabilities of .118, whereas the same statistic in population 2 yields an average deviation of .045. The latent classes in population 1 are less clearly defined because the conditional probabilities for the two latent classes are much closer as compared to population 2. Therefore, we would expect a greater error in the sample estimates of the conditional probabilities in population 1. The range of errors in samples for both populations was slightly more than double the sample with the lowest error, i.e., some samples gave less than half the estimation error as compared to other samples. These results are presented in Table VIII.

Addition of the supplementary tables B*C*D to the analysis resulted in an average 19% reduction in error in the first population and an average 16% reduction in error in the second population. Examination of individual samples indicates that there was

TABLE VII - Simulation results

Mean Absolute Deviations From Population Conditional Probabilities *

population	sample	complete table only	complete + B*C*D	complete + B*C*D + A*C*D	B*C*D + A*B*D
1	1	.156	.121	.118	.188
"	2	.120	.086	.076	.185
"	3	.076	.065	.058	.061
"	4	.098	.093	.118	.157
"	5	.106	.069	.084	.105
"	6	.068	.032	.053	.131
"	7	.123	.129	.087	.081
"	8	.133	.134	.066	.072
"	9	.168	.119	.118	.112
"	10	.132	.117	.081	.061
	mean	.118	.096	.086	.115
standard error		.0101	.0105	.0078	.0153
paired t-value**			3.59	4.15	0.16

* means are averages over the sixteen conditional probabilities of the absolute differences between the sample and population values.

** the pairing is between the "complete table only" column and the other three columns; a one-tailed t-value greater than 1.83 is statistically significant for the second and third columns, whereas a two-tailed t-value of 2.26 is significant for the last column at alpha levels of .05.

Table VII (continued)

Mean Absolute Deviations From Population Conditional Probabilities

population	sample	complete table only	complete + B*C*D	complete + B*C*D + A*C*D	B*C*D + A*B*D
2	1	.058	.044	.019	.032
"	2	.030	.021	.017	.034
"	3	.037	.038	.025	.040
"	4	.065	.056	.024	.060
"	5	.055	.025	.020	.052
"	6	.035	.039	.023	.026
"	7	.054	.042	.034	.058
"	8	.043	.033	.027	.048
"	9	.041	.048	.044	.059
"	10	.032	.037	.035	.056
	mean	.045	.038	.027	.047
standard error		.0038	.0033	.0027	.0040
paired t-value**			1.88	3.64	-0.34

* means are averages over the sixteen conditional probabilities of the absolute differences between the sample and population values.

** the pairing is between the "complete table only" column and the other three columns; a one-tailed t-value greater than 1.83 is statistically significant for the second and third columns, whereas a two-tailed t-value of 2.26 is significant for the last column at alph levels of .05.

virtually no gain in 3/10 samples in the first population and no gain (or slightly worse) in 4/10 samples in the second population. The rest of the samples showed substantial gain with one sample in each population reducing the error by >50%.

Including a second supplementary table, an A*C*D marginal table, further improved conditional probability estimates on average in both populations, although similar to the case of a single supplementary table, there was variation in the gain achieved. There was one case (population 1, sample 4) where the average conditional probability estimate was 20% worse than analysis of the complete table only. The average error reduction was 27% in the first population and 40% in the second population as compared to analyzing the complete table only.

Analyzing two supplementary tables together, B*C*D and A*C*D, in the absence of a complete table yielded, on average, similar results to analyzing the complete data table only.

The paired-t statistics presented in the table show that addition of supplementary tables to the complete table results in significant reductions in the deviation index. Analysis of the two supplementary tables in the absence of the complete table gives similar results to analysis of the complete table only.

IV. Generalized Linear Models

A "complete" population with a supplementary B*C*D population and a supplementary A*B*D populations, both missing at random with respect to variable D, were generated via a generalized linear model. The method of generating these frequencies is instructive

since Latent Class Analysis via GLM with and without supplementary tables is set up in an analogous fashion. The general methodology for models with missing data can be found in Rindskopf(1989).

First, a design matrix is constructed. Each column represents an effect in the Latent Class model. There are columns for an intercept, main effects for each of the observed and latent class effects, interactions between the latent classes and each of the observed effects, missing data, and missing data by main effect interactions. As usual, if there are two levels of each effect then the effect can be represented by a single column of ones and zeros or ones and negative ones. Otherwise, the number of columns for an effect is one less than the number of levels. Each of these effects is explained in more detail below.

1) Intercept: the magnitude of the intercept is related to the average cell size but is affected by other effects in the model. In the example to be presented 96 cells were generated with a total sample size of 10,000 by using an intercept of 4.3, which is approximately ln(10,000/96).

2) Main effects for observed variables: zeros were used in the example to generate equal marginal frequencies of 5000 for each of the two levels of the four observed variables. Generally any coeffient greater than 1 will grossly imbalance the frequences of levels within an observed variable, e.g., if all other effects were zero except for an intercept of 4 then a β of 1 for a two-level effect would yield frequencies of 20 for one level and 150 for another level.

3) Latent class effects: these effects are similar in nature to the observed effects except they serve to define the latent classes.

4) Interactions between observed and latent class effects: these are the effects that differentiate Latent Class Models from other models. They are related to the conditional probabilities of MLLSA. The direct relationship is that the GLM effect estimates are equal to:

(ln(crossproduct of conditional probabilities)) / 4 .

For example, if the conditional probabilities for variable A in latent class 1 were .9 for A_1, .1 for A_2 and for latent class 2 were .1 for A_1 and .9 for A_2 then the latent class by observed variable interaction effect would be:

$$(\ln(.9 \times .9 / .1 \times .1)) / 4 = 1.0986 .$$

5) Missing data effects: these effects indicate the relative magnitude of missing data. There will be a column for each type of missing data. For example, if there were a supplementary table where data was missing with respect to A and there were a second table where data were missing with respect to B then there would be a column for A and a separate column for B.

6) Missing data by observed variable interactions: these define effects due to MAR. For example, if an observed variable were MAR with respect to observed variable D then a missing by D interaction column would be inserted and the magnitude of this effect would determine the difference in

missingness of the missing variable relative to each level of D. A β=0.5 for a missing A variable MAR with respect to variable D led to a 73% missing of A for D_1 versus a 27% missing of A for D_2.

The Design Matrix together with β coefficients will produce predicted frequencies for the cells identified by the Design Matrix using the following equation:

$$\text{frequencies} = \exp(\ X * \beta)$$

The Design Matrix for the above problem is given in table VIII. There are a total of 64 rows in the design matrix with each row representing an individual cell frequency in the complete underlying table.

The frequencies obtained by multiplying the design matrix by the Beta coefficients are the individual cell frequencies in the complete underlying table. To obtain "observed" cell frequencies it is necessary to premultiply the frequency vector obtained above by a matrix, denoted as C, which is a matrix of 1's and 0's indicating which cells in the complete table are to be combined to form the observed table. The C matrix for this example is presented in table III. Each row in the C matrix indicates which cells are to be combined, and each section of rows indicates a different type of combining of data. For example, the first row says to combine cells $A_1B_1C_1D_1$ of the two latent classes that don't contain missing data; the second row says to combine cells $A_2B_1C_1D_1$ of the two latent classes that don't contain missing data, etc. Starting with the 17th row, cells are not only combined over latent classes but over variable A as well (note the four 1's in each of these rows).

Table VIII

Design Matrix for a 4 variable (2 levels per variable), 2 latent class, any observed variable MAR with respect to variable D

(X = Latent Class)

row	intcpt	A	B	C	D	X	A*X	B*X	C*X	D*X	Miss	Miss*D
1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	-1	1	1	1	1	-1	1	1	1	1	1
3	1	1	-1	1	1	1	1	-1	1	1	1	1
4	1	-1	-1	1	1	1	-1	-1	1	1	1	1
5	1	1	1	-1	1	1	1	1	-1	1	1	1
6	1	-1	1	-1	1	1	-1	1	-1	1	1	1
7	1	1	-1	-1	1	1	1	-1	-1	1	1	1
8	1	-1	-1	-1	1	1	-1	-1	-1	1	1	1
9	1	1	1	1	-1	1	1	1	1	-1	1	-1
10	1	-1	1	1	-1	1	-1	1	1	-1	1	-1
11	1	1	-1	1	-1	1	1	-1	1	-1	1	-1
12	1	-1	-1	1	-1	1	-1	-1	1	-1	1	-1
13	1	1	1	-1	-1	1	1	1	-1	-1	1	-1
14	1	-1	1	-1	-1	1	-1	1	-1	-1	1	-1
15	1	1	-1	-1	-1	1	1	-1	-1	-1	1	-1
16	1	-1	-1	-1	-1	1	-1	-1	-1	-1	1	-1

17-32 repeat rows 1-16 with -1 in X column modified X interaction columns

33-64 repeat rows 1-32 with -1 in Miss column and modified Miss*D interaction column

Table IX

"C" matrix for deriving observed cells counts for a MAR Latent Class model

row			
1	1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		0_{32}
2	0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		0_{32}
3	0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		0_{32}
4-15		etc.	
16	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1		0_{32}
17	0_{32}	1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
18	0_{32}	0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0	
19-23		etc.	
24	0_{32}	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1	

Analysis of a Latent Class model when given the observed frequencies involves construction of a Design Matrix, as above, construction of a "C" matrix, also as above, but requires an iterative procedure to derive estimates of the β parameters. A short SAS IML program (program 7 in Appendix I) was written to set up the observed tables.

An example of this method of analysis is provided as the program in Appendix II. Sample #1 from population #2 was analyzed with the B*C*D supplementary table. The natural logarithm of the crossproduct ratios for the conditional probabilities derived via MLLSA was four times the estimated β for that effect in GLM as shown in (4) above. In fact, initial values for GLM in this case were derived from conditional

probabilities obtained in the MLLSA procedure. Notice that there is no (missing by observed) variable column since the data were MCAR. For MAR data additional columns would be included. In general, poor initial values, i.e., initial values that are far from the final values, will cause the program to diverge without obtaining the final estimates that are sought.

V. MAR Tables

To illustrate the effects of MAR on MLLSA analyses two MAR populations were generated using the modelling techniques explained above. The first population had one B*C*D supplementary table where observed variable A was MAR with respect to variable D, i.e., the Design Matrix had a single column for missingness, and it had an additional column for the interaction between missingness and observed variable D. The following parameters were used:

β = 4.27	intercept
0	missingness - a "0" produces equal frequencies in the complete and supp. tables
0	latent class - a "0" produces equally sized latent classes
0,0,0,0	Observed variables A, B, C, D - produces equal marginals for non-missing variables
1.0986	these four coefficients are functions of conditional
0.6931	probabilities as indicated in (4)

-1.0986	above; these particular coefficients
-0.6931	are for population #2
0.5	this is the MAR parameter - for a missing A table MAR with respect to variable D led to a 73% loss of A_1-A_2 differentiation for D_1 versus a 27% loss of A_1-A_2 for D_2.

When the frequencies resulting from the model above were analyzed by MLLSA the following conditional probabilities were obtained:

Table X

Conditional probabilities computed via MLLSA for population #2 with variable A MAR with respect to D

	complete table		B*C*D table		
latent class:	1	2	1	2	
A_1	.10	90	1	1	
A_2	.90	.10	0	0	
B_1	.20	.80	.20	.80	
B_2	.80	.20	.80	.20	
C_1	.90	.10	.90	.10	
C_2	.10	.90	.10	.90	
D_1	.59	.08	.92	.41	(.80 .20
D_2	.41	.92	.08	.59	.20 .80 are theoretical values)

For variables A, B, and C the theoretical conditional probabilities for population #2 have been recovered, whereas the conditional probabilities for variable D are quite different from the theoretical values (as per the rules for analyzing MAR data, variable D's conditional probabilities are *not* restricted to be equal in the complete and supplementary tables). What has happened is that the A variable is more heavily collapsed with respect to D_1 relative to D_2 and therefore the D_1 probabilities in the complete table are diminished and the D_1 probabilities in the supplementary table are increased. This example also illustrates the theoretical recovery of A's conditional probabilities; it is the conditioning variable, i.e., D in this case, which is affected.

Another analysis was performed similar to the above except that two supplementary tables, B*C*D and A*C*D, were created. An additional missing column was added to the Design Matrix as well as a column of the interaction between this missing column and variable B. The same parameter vector as in the previous example was used except that β=0 was added for the additional missing column and β = -0.5 was added for the missing by variable D interaction column. Results are presented in Table XI.

Notice that for the supplementary tables the distortion in conditional probabilities is equal but in the opposite direction; this is a result of betas of equal magnitude but in the opposite direction (0.5, -0.5). In addition, the conditional probabilities for D in the complete table are correct due to the balancing of the betas for MAR. In general, none of the conditional probabilities in the complete or supplementary tables will be correct except in a special case such as we have here. However, the crossproduct of the conditional probabilities for an observed variable will be an unbiased estimator for

Table XI

Conditional probabilities computed via MLLSA for
population #2 with A and B MAR with respect to D

	complete table		B*C*D table		A*C*D table		
latent class:	1	2	1	2	1	2	
A_1	.10	90	1	1	.10	.90	
A_2	.90	.10	0	0	.90	.10	
B_1	.20	.80	.20	.80	1	1	
B_2	.80	.20	.80	.20	0	0	
C_1	.90	.10	.90	.10	.90	.10	
C_2	.10	.90	.10	.90	.10	.90	
D_1	.80	.20	.92	.41	.59	.08	(.80 .20
D_2	.20	.80	.08	.59	.41	.92	.20 .80 are theoretical values)

the crossproduct in the complete table (.8 x .8 / .2 x .2 = .92 x .59 / .08 x .41).

DISCUSSION

The techniques of incorporating marginal tables that have data missing with regard to one or more observed tables have been given and illustrated above. More complex situations could also be handled by these techniques, such as when: (a) there is a supplementary table that is MCAR and a second supplementary table that is MAR with respect to an observed variable; (b) there is a supplementary table that is MAR with respect to an interaction between two observed variables; (c)there is a supplementary table that is missing data with regard to two or more variables and is MAR with respect to one or more observed variables.

There are special situations where a set of variables are all considered important to discriminate among the latent classes but for reasons of cost or time or disclosure, it is not possible to obtain completely cross-classified data with respect to all variables. Instead, a series of marginal tables are obtained; in the extreme it might only be possible to have series of related tables that are cross-classified with respect to two variables. Ordinarily it would not be possible to perform Latent Class Analysis in such cases but with the techniques laid out above it would be possible. If one were to set up the conditional probability matrix for such a case involving 4 observed variables it would look like Table XII.

Table XII

Conditional probability configuration for a series of two variable supplementary tables

	A*B		A*C		A*D		B*C		B*D		C*D	
latent class:	1	2	1	2	1	2	1	2	1	2	1	2
A_1	a_1	$1-a_2$	a_1	$1-a_2$	a_1	$1-a_2$	1	1	1	1	1	1
A_2	$1-a_1$	a_2	$1-a_1$	a_2	$1-a_1$	a_2	0	0	0	0	0	0
B_1	b_1	$1-b_2$	1	1	1	1	b_1	$1-b_2$	b_1	$1-b_2$	1	1
B_2	$1-b_1$	b_2	0	0	0	0	$1-b_1$	b_2	$1-b_1$	b_2	0	0
C_1	1	1	c_1	$1-c_2$	1	1	c_1	$1-c_1$	1	1	c_1	$1-c_2$
C_2	0	0	$1-c_1$	c_2	0	0	$1-c_2$	c_2	0	0	$1-c_1$	c_2
D_1	1	1	1	1	d_1	$1-d_2$	1	1	d_1	$1-d_2$	d_1	$1-d_2$
D_2	0	0	0	0	1-d1	d_2	0	0	$1-d_1$	d_2	$1-d_1$	d_2

Assuming the missing data were MCAR then all the A_1's would be restricted to be equal, all the B_1's would be restricted to be equal etc. If the data were MAR the conditional probabilities would not be correct but if it could be assumed that the data were all derived from the same population the the discrimination among latent classes should be correct.

If Latent Class Analysis is performed without MAR supplementary tables that have been derived from the same sample, i.e, if a sample is collected and subsequently it is discovered that there is missing data on one variable the extent of which is dependent upon levels of another variable, then Latent Class Analysis on the complete table only will yield biased results. This can be seen by referring to Table X in the results section. Suppose say that we analyzed data only from the complete table; it is obvious that because of the distortion of the conditional probabilities for variable D we will not obtain an unbiased Latent Class Analysis. Incorporation of the supplementary MAR table will insure an unbiased analysis.

For MCAR we will obtain unbiased estimates whether we decide to use supplementary tables or not. But discarding the incomplete tables will diminish the statistical power of the procedure by effectively diminishing the sample size.

The value of incorporating supplementary tables in Latent Class Analysis was demonstrated by the simulations. In these simulations each suplementary table was equal in size to the complete table. Judging by the improvement in recovering conditional probabilities under these conditions it can be concluded that if there are supplemental tables with as many subjects as in the complete table then it will be well worth the effort to use this additional data to improve estimates. If the supplemental tables are much smaller than the complete table then it may not be worth the effort.

How much the analysis is helped by incorporation of supplementary tables is not only a function of relative sample size of the supplementary tables and degree of MAR, but is also a function of the relative discrimination of the variables in the supplementary table. For example, if we have the following population (Table XIII):

Table XIII

Discrimination of observed variables

latent class:	1	2
A_1	.9	.1
A_2	.1	.9
B_1	.1	.9
B_2	.9	.1
C_1	.45	.55
C_2	.55	.45
D_1	.55	.45
D_2	.45	.55

Inclusion of an A*B MCAR table would be of much greater benefit than including a C*D supplementary table because both variables A and B are highly discriminatory variables whereas C and D have do not discriminate well between the latent classes.

Both the published program MLLSA (Clogg, 1977) and the Maximum Likelihood program written for this dissertation (Appendix II) use the same algorithm and yield the same results if they are parameterized the same, i.e., the same conditional probabilities, latent class probabilities, convergence criterion. The difference relates to the ease of use and flexibility of the programs. When there are no restrictions the programs are about equivalent to use, with the exception that for SAS MLLSA one must enter an additional matrix of 1's and 0's indicating which rows are to be combined to form marginal counts. In practice this is easy to do. When there are restrictions Clogg's MLLSA program

has an easy method to indicate which probabilities are to be restricted; each element of a matrix represents a conditional probability and by use of a special numbering system it is easy to indicate the restrictions. The FORTRAN code then uses this matrix to equate conditional probabilities.

For SAS MLLSA it is necessary to construct the restrictions directly which consists of taking averages of weighted (by latent class probabilities) conditional probabilities and equating one conditional probability to another. In doing this one must reference specific elements of the row x column conditional probability matrix. If there are large numbers of restrictions then this is cumbersome and prone to error.

The major benefit of using SAS MLLSA is its flexibility. External datasets may be read directly into SAS MLLSA; with Goodman's MLLSA the data has to be inserted manually into each run, although for both programs the different runs may be "stacked" so that many analyses can be obtained in one execution of the programs. Basically, Clogg's program is a fixed program with fixed input/output which can be modified somewhat by a fixed set of options. The FORTRAN code is available to users who wish to modify the program but FORTRAN coding can be time consuming and error-ridden.

With SAS MLLSA the user can easily modify the program to get additional statistics, or fewer statistics, and can easily build upon the basic program with a little knowledge of SAS IML. SAS IML is much easier to learn and remember than FORTRAN and is a much more powerful language with many pre-programmed functions and capabilities.

Performing Latent Class Analysis using the Generalized Linear Models Approach has advantages and disadvantages compared with the MLLSA. Three disadvantages are: (1) Large matrices, the Design and "C" matrices, frequently have to be input which can be tedious and error-prone. SAS IML and other matrix oriented statistical packages do however contain some functions for generating matrices which have repetative components; this eases but does not erase the burden of specifying these matrices. (2) Convergence to a solution can be difficult if the initial values are not reasonably close to the final values. If one has a good feel for the data-structure then one is more likely to obtain reasonable start values as opposed to randomly choosing start values. The next paragraph summarizes how to obtain reasonable initial values. (3) The β parameters are not directly interpretable in the same way as MLLSA. For example, the conditional probabilities of MLLSA are not obtainable directly from GLM, although a function of these is obtainable, namely the cross-product ratio of the conditional probabilities.

Using the information on how the author constructed the simulation populations through GLM in the Results section one could often obtain reasonable start values, although convergence could still be a problem. Basic observations are: (1) the intercept is related to the natural logarithm of the average cell size. (2) Coefficients of "0" for other effects indicate equality between the levels being contrasted. (3) Coefficients different from zero for observed variable main effects, for latent class, and for missing data are related to the intercept (in this research any one of these effects parameterized to be greater than 1 greatly imbalanced the data with respect to that effect). (4) Coefficients for the latent class by observed variable interaction are directly related to the logarithm of the cross-product ratio of the conditional probabilities for that observed variable so that if one has good intuition as to what

the conditional probabilities should be then he can transform then directly into good estimates of Beta coefficients. (5) For MAR one should have a good guess as to the relative loss of variable X with respect to levels of variable Y and this can be transformed into an appropriate beta; for instance, a beta of 0.5 for the MAR column, as previously reported, resulted in 73% of variable A being collapsed for values of D_1 versus 27% of variable A being collapsed for values of D_2.

One advantage of GLM is the ability to isolate and estimate MAR parameters. From MLLSA it is not possible to get a direct estimate of MAR, although one could infer the degree of MAR by examining the relative differences in conditional probabilities for a complete and supplementary table for the conditioning variable. GLM, however, provides a direct estimate of MAR by including a missing by observed interaction variable (the MAR variable) column in the Design Matrix.

Another advantage of GLM is its ability to estimate more complex models than MLLSA. Latent Class Analysis is a very specific model, but it is conceivable that one may be working with a model that slightly more general than LCA, i.e., there may be some additional terms that are unaccounted for in LCA but can be parameterized in GLM. For instance, MLLSA cannot test for MCAR but by adding (missing by observed variable) columns in the GLM Design Matrix effects tests can be performed to determine if the missing data are MAR or MCAR.

MLLSA has advantages of relative ease of use and easily interpretable parameters in terms of the conditional and latent class probabilities. In addition, one only has to input initial values for conditional and latent class probabilities. For GLM, if there are many levels for a variable one must supply one less than the number of levels for

effect estimates and these estimates theoretically can be any real number, whereas in MLLSA one simply assigns proportions to the levels so they add up to one.

Both approaches, MLLSA and GLM, use the same model for performing basic Latent Class Analysis. They also both yield maximum likelihood estimates. The difference between the two is computational: MLLSA uses the EM algorithm whereas GLM uses an iterative weighted least squares algorithm.

Future work relating to these methods should include: (1) examples of application of these methods to unpublished datasets where the complete dataset is available. Currently published datasets only contain the completely observed cross-classifications. (2) Extention of these techniques to situations in which the data are MAR with respect to variables within latent classes, i.e., where it has been determined that "missingness" is related to the particular latent class an observation belongs to. (3) A study of the standard errors of the parameter estimate estimates as they relate to the proportion of partially-missing data, degree of MAR, and multinomial parameters (e.g., effect of a binomial parameter of .5 versus a parameter of .1).

In conclusion, the results of the research presented herein extend and illustrate the methods of Thompson and Baker (1981), Rubin and Little (1987), and Rindskopf (1989) relating to analysis of partially missing data specifically to the domain of Latent Class Analysis. The results indicate that the methods are relative easy to implement in either GLM or MLLSA and that if supplementary marginal tables that have frequencies equal to or greater that the "complete" table are not included in the analysis serious inefficiencies and/or biases can result. In addition, Latent Class analyses including all relevant variables have not been possible when each subject

was missing data on one or more variables; analysis including only partially cross-classified was demonstrated to not only be possible but to yield highly desirable results.

In addition, an entirely new program written in SAS IML is presented. This program is an alternative to Goodman's MLLSA FORTRAN program for those who have access to SAS IML and provides a greater degree of flexibility in input/output and modifications as compared to Goodman's MLLSA FORTRAN program.

APPENDIX I

Programs are minimally documented below; other documentation is contained in the program input/output listings.

Programs/output contained in this appendix are:

Program #1

This is a program input and output listing of MLLSA performed on the complete population #1 (10,000 observations). This illustrates use of the MLLSA program and verifies that the frequencies are those obtained by the unconditional latent class probabilities and conditional probabilities in table V.

Program #2

This is a program input and output listing of MLLSA performed on the complete population #1 plus two supplementary tables, the first consisting of population #1 collapsed over observed variable A and the second consisting of population #1 collapsed over observed variable B. This program illustrates the method of data input for the supplementary tables, the set-up of the additional latent classes to accomodate the supplementary tables, and the pattern of restrictions on the conditional probabilities. It also verifies that the collapsing was performed correctly in obtaining MCAR supplementary tables; this is evident in the "recapturing" of the conditional and latent class probabilities (within round-off error).

Program #3

This is the same as Program #1 except that the analysis is for population #2.

Program #4

This is the same as Program #2 except that the analysis is for population #2.

Program #5

This is a listing of the SAS IML program that created 30 multinomial samples of 100 observations apiece from each of the two populations.

Program #6

This is a listing of the SAS IML program that created the supplementary tables collapsed over observed variables A and B. Data input is the 30 multinomial samples created by program #5.

Program #7

This is a listing of the program and output of a SAS IML program that generates a complete table and two supplementary tables which are MAR with respect to variable D; the first supplementary table is missing variable A and the second is missing variable B. Program input is the Design Matrix "X", the "C" matrix which indicates which cells of the underlying table are combined to form the observed table, and the *Beta* coefficients. The model used is:

$$\text{observed cell frequencies} = C * \exp (X * Beta) .$$

Appendix 1 Program #1

```
MLLSA on Complete Population #1
 4 2          150  .00005 1                    1 2
 2 2 2 2
FREE
72 66 312 370 90 72 306 312 504 378 1512 1386 714 504 1890 1512
.3      .7
.4      .6      .6      .4
.1      .9      .3      .7
.6      .4      .4      .6
.3      .7      .1      .9
```

```
MLLSA on Complete Population #1
TOTAL OF INPUT, AND EXPECTED  10000.      0.
    72.     66.    312.    370.     90.     72.    306.    312.    504.    378.
  1512.   1386.    714.    504.   1890.   1512.

INPUT X'S  .3000000  .7000000

INPUT Y'S
 .4000000  .6000000  .6000000  .4000000
 .1000000  .9000000  .3000000  .7000000
 .6000000  .4000000  .4000000  .6000000
 .3000000  .7000000  .1000000  .9000000

MARGINALS FOR IND MODEL
1        5400.   4600.
2        2400.   7600.
3        4600.   5400.
4        1600.   8400.

CHI-SQUARE FOR IND MODEL  LR=     110.691  PRSN=    116.167

X RESTRICTIONS

Y RESTRICTIONS

DEVIATION    .11102230E-15 ITERATIONS     1
```

Appendix I Program #1

```
FIT
    72.00      66.00     312.00     370.00      90.00      72.00     306.00     312.00
   504.00     378.00    1512.00    1386.00     714.00     504.00    1890.00    1512.00

  CUMULATED CHISQUARE, LR=       .000  PEARSON=       .000

  CUMULATED CHISQUARE, LR=       .000  PEARSON=       .000

FINAL LR       .000 PRSN        .000 INDEX OF DISSIMILARITY      .0000

FINAL LATENT CLASS PROBABILITIES
 .3000000  .7000000

FINAL CONDITIONAL PROBABILITIES

LATENT CLASS=          1        2        3        4

1.        1.              .4000    .6000

1.        2.              .6000    .4000

2.        1.              .1000    .3000

2.        2.              .9000    .7000

3.        1.              .6000    .4000

3.        2.              .4000    .6000

4.        1.              .3000    .1000

4.        2.              .7000    .9000

NUMBER OF ESTIMATED PARAMETERS=      9   DEGREES OF FREEDOM IF IDENTIFIED=      6

COLUMN RANK=      9 DEGREES OF FREEDOM=      6

MLLSA on Complete Population #2
TOTAL OF INPUT, AND EXPECTED   9345.       0.
   144.     656.     306.    2594.     144.     194.     306.     306.     194.     144.
   656.    2594.     306.     656.     144.       1.

INPUT X'S  .1000000  .9000000

INPUT Y'S
 .2000000  .8000000  .8000000  .2000000
 .9000000  .1000000  .1000000  .9000000
 .8000000  .2000000  .2000000  .8000000
```

Appendix I Program #2

```
 Population #1   MCAR  with Two Supplementary Tables
 5 6          150          1            1        1 2
 2 2 2 2 3
FREE
72 66 312 370 90 72 306 312 504 378 1512 1386 714 504 1890 1512
138 0 682 0 162 0 618 0 882 0 2898 0 1218 0 3402 0
384 436 0 0 396 384 0 0 2016 1764 0 0 2604 2016 0 0
.3       .7       .3       .7       .3       .7
 .4       .6       .6       .4      1.0      0.0      1.0      0.0       .4       .6
 0.6       .4
 .1       .9       .3       .7       .1       .9       .3       .7      1.0      0.0
1.0      0.0
 .6       .4       .4       .6       .6       .4       .4       .6       .6       .4
 .4       .6
 .3       .7       .1       .9       .3       .7       .1       .9       .3       .7
 .1       .9
1.0      0.0      0.0      1.0      0.0      0.0      0.0      1.0      0.0      0.0
1.0      0.0      0.0      0.0      1.0      0.0      0.0      1.0
   2   0   3   0   1   0   1   0   2   0   3   0
   4   0   5   0   4   0   5   0   1   0   1   0
   6   0   7   0   6   0   7   0   6   0   7   0
   8   0   9   0   8   0   9   0   8   0   9   0
   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1  -1   1   1
```

Appendix I Program #2

```
1 Population #1   MCAR  with Two Supplementary Tables

-TOTAL OF INPUT, AND EXPECTED  30000.      0.
     72.     66.    312.    370.     90.     72.    306.    312.    504.    378.
   1512.   1386.    714.    504.   1890.   1512.
    138.      0.    682.      0.    162.      0.    618.      0.    882.      0.
   2898.      0.   1218.      0.   3402.      0.
    384.    436.      0.      0.    396.    384.      0.      0.   2016.   1764.
      0.      0.   2604.   2016.      0.      0.
0INPUT X'S  .3000000  .7000000  .3000000  .7000000  .3000000  .7000000
0INPUT Y'S
  .4000000  .6000000  .6000000  .4000000 1.0000000  .0000000 1.0000000  .0000000
  .4000000  .6000000  .6000000  .4000000
  .1000000  .9000000  .3000000  .7000000  .1000000  .9000000  .3000000  .7000000
 1.0000000  .0000000 1.0000000  .0000000
  .6000000  .4000000  .4000000  .6000000  .6000000  .4000000  .4000000  .6000000
  .6000000  .4000000  .4000000  .6000000
  .3000000  .7000000  .1000000  .9000000  .3000000  .7000000  .1000000  .9000000
  .3000000  .7000000  .1000000  .9000000
 1.0000000  .0000000  .0000000 1.0000000  .0000000  .0000000  .0000000 1.0000000
  .0000000  .0000000 1.0000000  .0000000  .0000000  .0000000 1.0000000  .0000000
  .0000000 1.0000000
0MARGINALS FOR IND MODEL
 1       20800.   9200.
 2       14800.  15200.
 3       13800.  16200.
 4        4800.  25200.
 5       10000.  10000.  10000.
0CHI-SQUARE FOR IND MODEL  LR=   29155.520  PRSN=  25098.340
0X RESTRICTIONS
0Y RESTRICTIONS
0LATENT CLASS =        1   2   3   4
 1.       1.          2   3   1   1   2   3
 1.       2.          0   0   0   0   0   0
 2.       1.          4   5   4   5   1   1
 2.       2.          0   0   0   0   0   0
 3.       1.          6   7   6   7   6   7
 3.       2.          0   0   0   0   0   0
 4.       1.          8   9   8   9   8   9
 4.       2.          0   0   0   0   0   0
 5.       1.          1   1   1   1   1   1
 5.       2.          1   1   1   1   1   1
 5.       3.          1   1   1   1   1   1
```

Appendix I Program #2

```
0DEVIATION     .38772823E-04 ITERATIONS     16
0FIT
    144.01     655.72     305.66    2594.85     656.27     143.99     194.10     305.60
    306.40     193.90     144.01     655.73    2593.14     306.34     656.28     143.99
    799.64        .00    2900.09        .00     800.35        .00     499.68        .00
    500.32        .00     799.65        .00    2899.91        .00     800.36        .00
    449.65    3250.37        .00        .00     850.42     449.57        .00        .00
    450.43     849.58        .00        .00    3249.62     450.35        .00        .00
0FINAL LR       .006 PRSN        .006 INDEX OF DISSIMILARITY      .0002
0FINAL LATENT CLASS PROBABILITIES
  .1666210  .1667123  .1666020  .1667314  .1666129  .1667204
0FINAL CONDITIONAL PROBABILITIES
0LATENT CLASS=          1         2         3         4

 1.        1.          .0999     .8999    1.0000    1.0000     .0999     .8999

 1.        2.          .9001     .1001     .0000     .0000     .9001     .1001

 2.        1.          .1999     .7999     .1999     .7999    1.0000    1.0000

 2.        2.          .8001     .2001     .8001     .2001     .0000     .0000

 3.        1.          .9001     .1001     .9001     .1001     .9001     .1001

 3.        2.          .0999     .8999     .0999     .8999     .0999     .8999

 4.        1.          .8001     .2001     .8001     .2001     .8001     .2001

 4.        2.          .1999     .7999     .1999     .7999     .1999     .7999

 5.        1.         1.0000    1.0000     .0000     .0000     .0000     .0000

 5.        2.          .0000     .0000    1.0000    1.0000     .0000     .0000

 5.        3.          .0000     .0000     .0000     .0000    1.0000    1.0000

NUMBER OF ESTIMATED PARAMETERS=    13   DEGREES OF FREEDOM IF IDENTIFIED=    34

COLUMN RANK=  13 DEGREES OF FREEDOM=    34
```

Appendix I Program #3

```
0FINAL LATENT CLASS PROBABILITIES
  .5000000   .5000000
0FINAL CONDITIONAL PROBABILITIES
0LATENT CLASS=          1        2        3        4

 1.         1.              .1000    .9000

 1.         2.              .9000    .1000

 2.         1.              .2000    .8000

 2.         2.              .8000    .2000

 3.         1.              .9000    .1000

 3.         2.              .1000    .9000

 4.         1.              .8000    .2000

 4.         2.              .2000    .8000

NUMBER OF ESTIMATED PARAMETERS=      9   DEGREES OF FREEDOM IF IDENTIFIED=      6

COLUMN RANK=      9 DEGREES OF FREEDOM=      6
```

Appendix I Program #3

```
ILLSA on Complete Population #2
4 2          150  .00005 1                    1 2
2 2 2 2
REE
44 656 306 2594 656 144 194 306 306 194 144 656 2594 306 656 144
5       .5
1       .9       .9       .1
2       .8       .8       .2
9       .1       .1       .9
8       .2       .2       .8
```

```
1MLLSA on Complete Population #2

-TOTAL OF INPUT, AND EXPECTED  10000.       0.
    144.    656.    306.   2594.    656.    144.    194.    306.    306.    194.
    144.    656.   2594.    306.    656.    144.
0INPUT X'S  .5000000  .5000000
0INPUT Y'S
  .1000000  .9000000  .9000000  .1000000
  .2000000  .8000000  .8000000  .2000000
  .9000000  .1000000  .1000000  .9000000
  .8000000  .2000000  .2000000  .8000000
0MARGINALS FOR IND MODEL
 1         5000.   5000.
 2         5000.   5000.
 3         5000.   5000.
 4         5000.   5000.
0CHI-SQUARE FOR IND MODEL  LR=   10674.040  PRSN=  15138.840
0X RESTRICTIONS
0Y RESTRICTIONS
0DEVIATION     .00000000E+00 ITERATIONS      1
0FIT
    144.00    656.00    306.00   2594.00    656.00    144.00    194.00    306.00
    306.00    194.00    144.00    656.00   2594.00    306.00    656.00    144.00
0  CUMULATED CHISQUARE, LR=       .000  PEARSON=       .000
0  CUMULATED CHISQUARE, LR=       .000  PEARSON=       .000
0FINAL LR      .000 PRSN       .000 INDEX OF DISSIMILARITY    .0000
```

Appendix I Program #4

```
  Population #2     MCAR  with Two Supplementary Tables
 5 6          150 .0005   1            1       1 2
 2 2 2 2 3
FREE
144 656 306 2594 656 144 194 306 306 194 144 656 2594 306 656 144
800 0 2900 0 800 0 500 0 500 0 800 0 2900 0 800 0
450 3250 0 0 850 450 0 0 450 850 0 0 3250 450 0 0
.15      .20      .15      .20      .15      .15
 .1       .9       .9       .1      1.0      0.0      1.0      0.0       .1       .9
 0.9       .1
 .2       .8       .8       .2       .2       .8       .8       .2      1.0      0.0
1.0      0.0
 .9       .1       .1       .9       .9       .1       .1       .9       .9       .1
 .1       .9
 .8       .2       .2       .8       .8       .2       .2       .8       .8       .2
 .2       .8
1.0      0.0      0.0      1.0      0.0      0.0      0.0      1.0      0.0      0.0
1.0      0.0      0.0      0.0      1.0      0.0      0.0      1.0
   2   0   3   0   1   0   1   0   2   0   3   0
   4   0   5   0   4   0   5   0   1   0   1   0
   6   0   7   0   6   0   7   0   6   0   7   0
   8   0   9   0   8   0   9   0   8   0   9   0
   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
```

Appendix I Program #4

```
 Population #2     MCAR  with Two Supplementary Tables

TOTAL OF INPUT, AND EXPECTED  30000.       0.
    144.     656.     306.    2594.     656.     144.     194.     306.     306.     194.
    144.     656.    2594.     306.     656.     144.
    800.       0.    2900.       0.     800.       0.     500.       0.     500.       0.
    800.       0.    2900.       0.     800.       0.
    450.    3250.       0.       0.     850.     450.       0.       0.     450.     850.
      0.       0.    3250.     450.       0.       0.
INPUT X'S   .1500000   .2000000   .1500000   .2000000   .1500000   .1500000
INPUT Y'S
  .1000000   .9000000   .9000000   .1000000 1.0000000   .0000000 1.0000000   .0000000
  .1000000   .9000000   .9000000   .1000000
  .2000000   .8000000   .8000000   .2000000   .2000000   .8000000   .8000000   .2000000
 1.0000000   .0000000 1.0000000   .0000000
  .9000000   .1000000   .1000000   .9000000   .9000000   .1000000   .1000000   .9000000
  .9000000   .1000000   .1000000   .9000000
  .8000000   .2000000   .2000000   .8000000   .8000000   .2000000   .2000000   .8000000
  .8000000   .2000000   .2000000   .8000000
 1.0000000   .0000000   .0000000 1.0000000   .0000000   .0000000   .0000000 1.0000000
  .0000000   .0000000 1.0000000   .0000000   .0000000   .0000000 1.0000000   .0000000
  .0000000 1.0000000
MARGINALS FOR IND MODEL
1        20000.  10000.
2        20000.  10000.
3        15000.  15000.
4        15000.  15000.
5        10000.  10000.  10000.
CHI-SQUARE FOR IND MODEL  LR=   44109.780  PRSN=  63228.360
X RESTRICTIONS
Y RESTRICTIONS
LATENT CLASS =          1   2   3   4
1.          1.          2   3   1   1   2   3.
1.          2.          0   0   0   0   0   0
2.          1.          4   5   4   5   1   1
2.          2.          0   0   0   0   0   0
3.          1.          6   7   6   7   6   7.
3.          2.          0   0   0   0   0   0
4.          1.          8   9   8   9   8   9
4.          2.          0   0   0   0   0   0
5.          1.          1   1   1   1   1   1
5.          2.          1   1   1   1   1   1
5.          3.          1   1   1   1   1   1
```

Appendix I Program #4

```
DEVIATION     .11102230E-15 ITERATIONS.     2
FIT.
    72.00     66.00    312.00    370.00     90.00     72.00    306.00    312.00
   504.00    378.00   1512.00   1386.00    714.00    504.00   1890.00   1512.00
   138.00       .00    682.00       .00    162.00       .00    618.00       .00
   882.00       .00   2898.00       .00   1218.00       .00   3402.00       .00
   384.00    436.00       .00       .00    396.00    384.00       .00       .00
  2016.00   1764.00       .00       .00   2604.00   2016.00       .00       .00
 FINAL LR       .000 PRSN        .000 INDEX OF DISSIMILARITY      .0000
 FINAL LATENT CLASS PROBABILITIES
  .1000000  .2333333  .1000000  .2333333  .1000000  .2333333
 FINAL CONDITIONAL PROBABILITIES
 LATENT CLASS=          1         2         3         4

 1.        1.          .4000     .6000    1.0000    1.0000     .4000     .6000

 1.        2.          .6000     .4000     .0000     .0000     .6000     .4000

 2.        1.          .1000     .3000     .1000     .3000    1.0000    1.0000

 2.        2.          .9000     .7000     .9000     .7000     .0000     .0000

 3.        1.          .6000     .4000     .6000     .4000     .6000     .4000

 3.        2.          .4000     .6000     .4000     .6000     .4000     .6000

 4.        1.          .3000     .1000     .3000     .1000     .3000     .1000

 4.        2.          .7000     .9000     .7000     .9000     .7000     .9000

 5.        1.         1.0000    1.0000     .0000     .0000     .0000     .0000

 5.        2.          .0000     .0000    1.0000    1.0000     .0000     .0000

 5.        3.          .0000     .0000     .0000     .0000    1.0000    1.0000

 NUMBER OF ESTIMATED PARAMETERS=     13    DEGREES OF FREEDOM IF IDENTIFIED=     34

 COLUMN RANK=     13 DEGREES OF FREEDOM=     34
```

```
options lt linesize=80;
data a;
  input cell_ID  prob @@;
*prob=prob/10000;

************NOTE: this is population #2***********;
/*
cards;
1111 144
2111 656
1211 306
2211 2594
1121 656
2121 144
1221 194
2221 306
1112 306
2112 194
1212 144
2212 656
1122 2594
2122 306
1222 656
2222 144
*/
/*   this is the 1st population obtained from population.sas */
cards;
1 72.0000
2 66.0000
3 312.0000
4 370.0000
5 90.0000
6 72.0000
7 306.0000
8 312.0000
9 504.0000
10 378.0000
11 1512.0
12 1386.0
13 714.0000
14 504.0000
15 1890.0
16 1512.0
;
proc print;
data a; set a;
cumprob+prob;
lcumprob=lag(cumprob);
if _n_=1 then lcumprob=0;
proc print;

proc iml;
 use a;
  read all;

num_samp=100;
totn=10000;
numcells=16;
prob=prob/10000;
  start iter;
```

Appendix I Program #5

```
do num=1 to 30;

if num=1 then seed=151;
if num=2 then seed=333;
if num=3 then seed=127;
if num=4 then seed=685;
if num=5 then seed=937;
if num=6 then seed=489;
if num=7 then seed=477;
if num=8 then seed=829;
if num=9 then seed=371;
if num=10 then seed=455;
if num=11 then seed=171;
if num=12 then seed=373;
if num=13 then seed=117;
if num=14 then seed=695;
if num=15 then seed=967;
if num=16 then seed=439;
if num=17 then seed=467;
if num=18 then seed=859;
if num=19 then seed=361;
if num=20 then seed=415;
if num=21 then seed=551;
if num=22 then seed=633;
if num=23 then seed=227;
if num=24 then seed=985;
if num=25 then seed=337;
if num=26 then seed=189;
if num=27 then seed=777;
if num=28 then seed=729;
if num=29 then seed=171;
if num=30 then seed=755;

cell_num={0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0};
do i=1 to num_samp;
pick=uniform(seed)*totn;

do j=1 to numcells;
if lcumprob(|j,|) < pick  then do;
if  pick < cumprob(|j,|) then do;
   cell_num(|j,|)=cell_num(|j,|)+1;

   end;
  end;
 end;
end;
print "100 observation sample from population #1 " seed;
print "Cell ID / Number in this cell / theoretical probability";
print cell_ID cell_num prob;
totsamp=cell_num(|+,|);
print totsamp;

end;
finish;
run iter;
stop;
quit;
```

```
filename in '[stewart.research.pops]mcar_pop_1.dat';
options lt linesize=80;
data table1;
*  length dum1 $5. dum2 $6. dum3 $5.;
  drop dum1 dum2 dum3;
  infile in;
  input dum1  dum2 dum3   freq1  ;


/* input observed frequency table as a column vector in the following order:

              P1111..., P2111..., P1211..., P2211...

       e.g., vary the first subscript the fastest and the last the slowest */
title 'POPULATION 1 ';
data table; set table1;
  if 1 le _n_ le 48;
;
proc iml;
use table ;
read all;

comp=freq1(|{1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16},|);

a1=freq1(|{17 19 21 23 25 27 29 31},|);
a2=freq1(|{18 20 22 24 26 28 30 32},|);
b1=freq1(|{33 34 37 38 41 42 45 46},|);
b2=freq1(|{35 36 39 40 43 44 47 48},|);

a=a1+a2;

b=b1+b2;

z=0;

zz={0,0};

freq=comp//a(|1,|)//z//a(|2,|)//z//a(|3,|)//z//a(|4,|)//z//a(|5,|)//z
//a(|6,|)//z//a(|7,|)//z//a(|8,|)//z//b(|{1 2},|)//zz//b(|{3 4},|)//zz//
b(|{5 6},|)//zz//b(|{7 8},|)//zz;


print freq;

n=freq(|+,|);

stop;
quit;


data table; set table1;
  if 49 le _n_ le 96;
;
proc iml;
use table ;
read all;

comp=freq1(|{1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16},|);
```

```
a1=freq1(|{17 19 21 23 25 27 29 31},|);
a2=freq1(|{18 20 22 24 26 28 30 32},|);
b1=freq1(|{33 34 37 38 41 42 45 46},|);
b2=freq1(|{35 36 39 40 43 44 47 48},|);

a=a1+a2;

b=b1+b2;

z=0;

zz={0,0};

freq=comp//a(|1,|)//z//a(|2,|)//z//a(|3,|)//z//a(|4,|)//z//a(|5,|)//z
//a(|6,|)//z//a(|7,|)//z//a(|8,|)//z//b(|{1 2},|)//zz//b(|{3 4},|)//zz//
b(|{5 6},|)//zz//b(|{7 8},|)//zz;


print freq;

n=freq(|+,|);

stop;
quit;


data table;set table1;
  if 97 le _n_ le 144;
;
proc iml;
use table ;
read all;

comp=freq1(|{1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16},|);

a1=freq1(|{17 19 21 23 25 27 29 31},|);
a2=freq1(|{18 20 22 24 26 28 30 32},|);
b1=freq1(|{33 34 37 38 41 42 45 46},|);
b2=freq1(|{35 36 39 40 43 44 47 48},|);

a=a1+a2;

b=b1+b2;

z=0;

zz={0,0};

freq=comp//a(|1,|)//z//a(|2,|)//z//a(|3,|)//z//a(|4,|)//z//a(|5,|)//z
//a(|6,|)//z//a(|7,|)//z//a(|8,|)//z//b(|{1 2},|)//zz//b(|{3 4},|)//zz//
b(|{5 6},|)//zz//b(|{7 8},|)//zz;


print freq;

n=freq(|+,|);

stop;
```

Appendix I Program #7

```
options lt linesize=80;
proc iml;

******** Statements in this section generate the design matrix *********;

 gen=  {   1   1   1   1   1,
           1  -1   1   1   1,
           1   1  -1   1   1,
           1  -1  -1   1   1,
           1   1   1  -1   1,
           1  -1   1  -1   1,
           1   1  -1  -1   1,
           1  -1  -1  -1   1,
           1   1   1   1  -1,
           1  -1   1   1  -1,
           1   1  -1   1  -1,
           1  -1  -1   1  -1,
           1   1   1  -1  -1,
           1  -1   1  -1  -1,
           1   1  -1  -1  -1,
           1  -1  -1  -1  -1,
          -1   1   1   1   1,
          -1  -1   1   1   1,
          -1   1  -1   1   1,
          -1  -1  -1   1   1,
          -1   1   1  -1   1,
          -1  -1   1  -1   1,
          -1   1  -1  -1   1,
          -1  -1  -1  -1   1,
          -1   1   1   1  -1,
          -1  -1   1   1  -1,
          -1   1  -1   1  -1,
          -1  -1  -1   1  -1,
          -1   1   1  -1  -1,
          -1  -1   1  -1  -1,
          -1   1  -1  -1  -1,
          -1  -1  -1  -1  -1 };

  misa=j(32,1,-1);
  misp=j(32,1,1);
  misn=j(32,1,0);
   mis=(misa//misp//misn)||(misa//misn//misp);

xx=j(96,1,1)||mis||repeat(gen,3,1);


xx=xx||j(96,6,0);


    xx(|,9|)=xx(|,4|)#xx(|,5|);
    xx(|,10|)=xx(|,4|)#xx(|,6|);
    xx(|,11|)=xx(|,4|)#xx(|,7|);
    xx(|,12|)=xx(|,4|)#xx(|,8|);


        /* MAR wrt D */
    xx(|,13|)=xx(|,2|)#xx(|,8|);
    xx(|,14|)=xx(|,3|)#xx(|,8|);
```

Appendix I Program #7

```
x=xx(|,{1 2 3 4 5 6 7 8 9 10 11 12 13 14 }|);

*************** End of Design Matrix generation ******************;

colms={"Int" "Miss A" "Miss B" "Class" "A" "B" "C" "D" "CL*A" "CL*B" "CL*C"
       "CL*D" "Miss A * D"  "Miss B * D"};

print "Design Matrix" ,
   x(|colname=colms format=4.0|);

******* The C matrix below indicates which rows in the
            Design Matrix are to be Combined  **********;

    /* Modelling Missing A and Missing B  Both MAR wrt D  */
 C = {1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
```

Appendix I Program #7

```
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.
0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.
0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0.
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Appendix I Program #7

```
print missing;


/* these were used to generate 1st pop */
b1={5.0186,-.423649,-.761131,0,-.761131,0,.3375,.2027,-.337482,-.202733};


/* these were used to generate MAR with respect to D for 2nd pop */

b1={4.2767, 0.0, 0.0 ,0,.3,-.2,.1,-.3,0.6931,1.0986,
     -.693147, -1.0986,  .5, -.5 };


print "these are the initial values" b1;


x=xx(|,{1 2 3 4 5 6 7 8 9 10 11 12 13 14 }|);


    eta = X*B1;

     H1 = exp(eta);

total=h1(|+,|);

h1=h1*(10000/total);

print "H1 values are the frequencies in the cells of the underlying table";
print h1;

 expctd=   c*h1;

rel_miss=missing'*h1;

print "Expected values are what is actually observed" ;
print expctd;

print "Relative missing data" rel_miss;
print "Total number of frequencies" total;

quit;
```

Appendix I Program #7

SAS 11:01 TUESDAY, OCTOBER 31, 1989 1

Design Matrix

X	Int	Miss A	Miss B	Clas s	A	B	C	D	CL*A	CL*B	CL*C	CL*D	Miss A * D	Miss B * D
ROW1	1	-1	-1	1	1	1	1	1	1	1	1	1	-1	-1
ROW2	1	-1	-1	1	-1	1	1	1	-1	1	1	1	-1	-1
ROW3	1	-1	-1	1	1	-1	1	1	1	-1	1	1	-1	-1
ROW4	1	-1	-1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
ROW5	1	-1	-1	1	1	1	-1	1	1	1	-1	1	-1	-1
ROW6	1	-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	-1
ROW7	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	-1
ROW8	1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	1	-1	-1
ROW9	1	-1	-1	1	1	1	1	-1	1	1	1	-1	1	1
ROW10	1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	1	1
ROW11	1	-1	-1	1	1	-1	1	-1	1	-1	1	-1	1	1
ROW12	1	-1	-1	1	-1	-1	1	-1	-1	-1	1	-1	1	1
ROW13	1	-1	-1	1	1	1	-1	-1	1	1	-1	-1	1	1
ROW14	1	-1	-1	1	-1	1	-1	-1	-1	1	-1	-1	1	1
ROW15	1	-1	-1	1	1	-1	-1	-1	1	-1	-1	-1	1	1
ROW16	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1
ROW17	1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	-1	-1
ROW18	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	-1
ROW19	1	-1	-1	-1	1	-1	1	1	-1	1	-1	-1	-1	-1
ROW20	1	-1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1
ROW21	1	-1	-1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1
ROW22	1	-1	-1	-1	-1	1	-1	1	1	-1	1	-1	-1	-1
ROW23	1	-1	-1	-1	1	-1	-1	1	-1	1	1	-1	-1	-1
ROW24	1	-1	-1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1
ROW25	1	-1	-1	-1	1	1	1	-1	-1	-1	-1	1	1	1
ROW26	1	-1	-1	-1	-1	1	1	-1	1	-1	-1	1	1	1
ROW27	1	-1	-1	-1	1	-1	1	-1	-1	1	-1	1	1	1
ROW28	1	-1	-1	-1	-1	-1	1	-1	1	1	-1	1	1	1
ROW29	1	-1	-1	-1	1	1	-1	-1	-1	-1	1	1	1	1
ROW30	1	-1	-1	-1	-1	1	-1	-1	1	-1	1	1	1	1
ROW31	1	-1	-1	-1	1	-1	-1	-1	-1	1	1	1	1	1
ROW32	1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1
ROW33	1	1	0	1	1	1	1	1	1	1	1	1	1	0
ROW34	1	1	0	1	-1	1	1	1	-1	1	1	1	1	0
ROW35	1	1	0	1	1	-1	1	1	1	-1	1	1	1	0
ROW36	1	1	0	1	-1	-1	1	1	-1	-1	1	1	1	0
ROW37	1	1	0	1	1	1	-1	1	1	1	-1	1	1	0
ROW38	1	1	0	1	-1	1	-1	1	-1	1	-1	1	1	0
ROW39	1	1	0	1	1	-1	-1	1	1	-1	-1	1	1	0
ROW40	1	1	0	1	-1	-1	-1	1	-1	-1	-1	1	1	0
ROW41	1	1	0	1	1	1	1	-1	1	1	1	-1	-1	0
ROW42	1	1	0	1	-1	1	1	-1	-1	1	1	-1	-1	0
ROW43	1	1	0	1	1	-1	1	-1	1	-1	1	-1	-1	0
ROW44	1	1	0	1	-1	-1	1	-1	-1	-1	1	-1	-1	0
ROW45	1	1	0	1	1	1	-1	-1	1	1	-1	-1	-1	0
ROW46	1	1	0	1	-1	1	-1	-1	-1	1	-1	-1	-1	0
ROW47	1	1	0	1	1	-1	-1	-1	1	-1	-1	-1	-1	0
ROW48	1	1	0	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0
ROW49	1	1	0	-1	1	1	1	1	-1	-1	-1	-1	1	0
ROW50	1	1	0	-1	-1	1	1	1	1	-1	-1	-1	1	0
ROW51	1	1	0	-1	1	-1	1	1	-1	1	-1	-1	1	0
ROW52	1	1	0	-1	-1	-1	1	1	1	1	-1	-1	1	0
ROW53	1	1	0	-1	1	1	-1	1	-1	-1	1	-1	1	0

Appendix I Program #7

ROW54	1	1	0	-1	-1	1	-1	1	1	-1	1	-1	1	0
ROW55	1	1	0	-1	1	-1	-1	1	-1	1	1	-1	1	0
ROW56	1	1	0	-1	-1	-1	-1	1	1	1	1	-1	1	0
ROW57	1	1	0	-1	1	1	1	-1	-1	-1	-1	1	-1	0
ROW58	1	1	0	-1	-1	1	1	-1	1	-1	-1	1	-1	0
ROW59	1	1	0	-1	1	-1	1	-1	-1	1	-1	1	-1	0
ROW60	1	1	0	-1	-1	-1	1	-1	1	1	-1	1	-1	0
ROW61	1	1	0	-1	1	1	-1	-1	-1	-1	1	1	-1	0
ROW62	1	1	0	-1	-1	1	-1	-1	1	-1	1	1	-1	0
ROW63	1	1	0	-1	1	-1	-1	-1	-1	1	1	1	-1	0
ROW64	1	1	0	-1	-1	-1	-1	-1	1	1	1	1	-1	0
ROW65	1	0	1	1	1	1	1	1	1	1	1	1	0	1
ROW66	1	0	1	1	-1	1	1	1	-1	1	1	1	0	1
ROW67	1	0	1	1	1	-1	1	1	1	-1	1	1	0	1
ROW68	1	0	1	1	-1	-1	1	1	-1	-1	1	1	0	1
ROW69	1	0	1	1	1	1	-1	1	1	1	-1	1	0	1
ROW70	1	0	1	1	-1	1	-1	1	-1	1	-1	1	0	1
ROW71	1	0	1	1	1	-1	-1	1	1	-1	-1	1	0	1
ROW72	1	0	1	1	-1	-1	-1	1	-1	-1	-1	1	0	1
ROW73	1	0	1	1	1	1	1	-1	1	1	1	-1	0	-1
ROW74	1	0	1	1	-1	1	1	-1	-1	1	1	-1	0	-1
ROW75	1	0	1	1	1	-1	1	-1	1	-1	1	-1	0	-1
ROW76	1	0	1	1	-1	-1	1	-1	-1	-1	1	-1	0	-1
ROW77	1	0	1	1	1	1	-1	-1	1	1	-1	-1	0	-1
ROW78	1	0	1	1	-1	1	-1	-1	-1	1	-1	-1	0	-1
ROW79	1	0	1	1	1	-1	-1	-1	1	-1	-1	-1	0	-1
ROW80	1	0	1	1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1
ROW81	1	0	1	-1	1	1	1	-1	-1	-1	-1	-1	0	1
ROW82	1	0	1	-1	-1	1	1	1	1	-1	-1	-1	0	1
ROW83	1	0	1	-1	1	-1	1	1	-1	1	-1	-1	0	1
ROW84	1	0	1	-1	-1	-1	1	1	1	1	-1	-1	0	1
ROW85	1	0	1	-1	1	1	-1	1	-1	-1	1	-1	0	1
ROW86	1	0	1	-1	-1	1	-1	1	1	-1	1	-1	0	1
ROW87	1	0	1	-1	1	-1	-1	1	-1	1	1	-1	0	1
ROW88	1	0	1	-1	-1	-1	-1	1	1	1	1	-1	0	1
ROW89	1	0	1	-1	1	1	1	-1	-1	-1	-1	1	0	-1
ROW90	1	0	1	-1	-1	1	1	-1	1	-1	-1	1	0	-1
ROW91	1	0	1	-1	1	-1	1	-1	-1	1	-1	1	0	-1
ROW92	1	0	1	-1	-1	-1	1	-1	1	1	-1	1	0	-1
ROW93	1	0	1	-1	1	1	-1	-1	-1	-1	1	1	0	-1
ROW94	1	0	1	-1	-1	1	-1	-1	1	-1	1	1	0	-1
ROW95	1	0	1	-1	1	-1	-1	-1	-1	1	1	1	0	-1
ROW96	1	0	1	-1	-1	-1	-1	-1	1	1	1	1	0	-1

MISSING	COL1	COL2	COL3
ROW1	1.0000	0	0
ROW2	1.0000	0	0
ROW3	1.0000	0	0
ROW4	1.0000	0	0
ROW5	1.0000	0	0
ROW6	1.0000	0	0
ROW7	1.0000	0	0
ROW8	1.0000	0	0
ROW9	1.0000	0	0

Appendix I Program #7

ROW10	1.0000	0	0
ROW11	1.0000	0	0
ROW12	1.0000	0	0
ROW13	1.0000	0	0
ROW14	1.0000	0	0
ROW15	1.0000	0	0
ROW16	1.0000	0	0
ROW17	1.0000	0	0
ROW18	1.0000	0	0
ROW19	1.0000	0	0
ROW20	1.0000	0	0
ROW21	1.0000	0	0
ROW22	1.0000	0	0
ROW23	1.0000	0	0
ROW24	1.0000	0	0
ROW25	1.0000	0	0
ROW26	1.0000	0	0
ROW27	1.0000	0	0
ROW28	1.0000	0	0
ROW29	1.0000	0	0
ROW30	1.0000	0	0
ROW31	1.0000	0	0
ROW32	1.0000	0	0
ROW33	0	1.0000	0
ROW34	0	1.0000	0
ROW35	0	1.0000	0
ROW36	0	1.0000	0
ROW37	0	1.0000	0
ROW38	0	1.0000	0
ROW39	0	1.0000	0
ROW40	0	1.0000	0
ROW41	0	1.0000	0
ROW42	0	1.0000	0
ROW43	0	1.0000	0
ROW44	0	1.0000	0
ROW45	0	1.0000	0
ROW46	0	1.0000	0
ROW47	0	1.0000	0
ROW48	0	1.0000	0
ROW49	0	1.0000	0
ROW50	0	1.0000	0
ROW51	0	1.0000	0
ROW52	0	1.0000	0
ROW53	0	1.0000	0
ROW54	0	1.0000	0
ROW55	0	1.0000	0
ROW56	0	1.0000	0
ROW57	0	1.0000	0
ROW58	0	1.0000	0
ROW59	0	1.0000	0
ROW60	0	1.0000	0
ROW61	0	1.0000	0
ROW62	0	1.0000	0
ROW63	0	1.0000	0
ROW64	0	1.0000	0
ROW65	0	0	1.0000
ROW66	0	0	1.0000
ROW67	0	0	1.0000

Appendix I Program #7

ROW68	0	0	1.0000
ROW69	0	0	1.0000
ROW70	0	0	1.0000
ROW71	0	0	1.0000
ROW72	0	0	1.0000
ROW73	0	0	1.0000
ROW74	0	0	1.0000
ROW75	0	0	1.0000
ROW76	0	0	1.0000
ROW77	0	0	1.0000
ROW78	0	0	1.0000
ROW79	0	0	1.0000
ROW80	0	0	1.0000
ROW81	0	0	1.0000
ROW82	0	0	1.0000
ROW83	0	0	1.0000
ROW84	0	0	1.0000
ROW85	0	0	1.0000
ROW86	0	0	1.0000
ROW87	0	0	1.0000
ROW88	0	0	1.0000
ROW89	0	0	1.0000
ROW90	0	0	1.0000
ROW91	0	0	1.0000
ROW92	0	0	1.0000
ROW93	0	0	1.0000
ROW94	0	0	1.0000
ROW95	0	0	1.0000
ROW96	0	0	1.0000

these are the initial values

B1	COL1
ROW1	4.2767
ROW2	0
ROW3	0
ROW4	0
ROW5	0.3000
ROW6	-0.2
ROW7	0.1000
ROW8	-0.3
ROW9	0.6931
ROW10	1.0986
ROW11	-.693147
ROW12	-1.0986
ROW13	0.5000
ROW14	-0.5

H1 values are the frequencies in the cells of the underlying table

H1	COL1
ROW1	18.5753
ROW2	2.5488
ROW3	3.0791

Appendix I Program #7

ROW4	0.4225
ROW5	60.8325
ROW6	8.3472
ROW7	10.0837
ROW8	1.3837
ROW9	304.6096
ROW10	41.7973
ROW11	50.4928
ROW12	6.9284
ROW13	997.5727
ROW14	136.8828
ROW15	165.3600
ROW16	22.6900
ROW17	18.5770
ROW18	40.7773
ROW19	249.4167
ROW20	547.4795
ROW21	3.8024
ROW22	8.3464
ROW23	51.0513
ROW24	112.0596
ROW25	3.7612
ROW26	8.2559
ROW27	50.4976
ROW28	110.8441
ROW29	0.7698
ROW30	1.6898
ROW31	10.3360
ROW32	22.6879
ROW33	30.6254
ROW34	4.2023
ROW35	5.0765
ROW36	0.6966
ROW37	100.2959
ROW38	13.7622
ROW39	16.6253
ROW40	2.2813
ROW41	184.7551
ROW42	25.3513
ROW43	30.6254
ROW44	4.2023
ROW45	605.0584
ROW46	83.0236
ROW47	100.2959
ROW48	13.7622
ROW49	30.6283
ROW50	67.2304
ROW51	411.2186
ROW52	902.6411
ROW53	6.2691
ROW54	13.7609
ROW55	84.1694
ROW56	184.7551
ROW57	2.2813
ROW58	5.0074
ROW59	30.6283
ROW60	67.2304
ROW61	0.4669

Appendix I Program #7

ROW62	1.0249
ROW63	6.2691
ROW64	13.7609
ROW65	11.2665
ROW66	1.5459
ROW67	1.8676
ROW68	0.2563
ROW69	36.8968
ROW70	5.0628
ROW71	6.1161
ROW72	0.8392
ROW73	502.2163
ROW74	68.9120
ROW75	83.2486
ROW76	11.4230
ROW77	1644.7
ROW78	225.6816
ROW79	272.6326
ROW80	37.4095
ROW81	11.2675
ROW82	24.7327
ROW83	151.2789
ROW84	332.0631
ROW85	2.3063
ROW86	5.0624
ROW87	30.9642
ROW88	67.9676
ROW89	6.2011
ROW90	13.6116
ROW91	83.2564
ROW92	182.7511
ROW93	1.2693
ROW94	2.7861
ROW95	17.0411
ROW96	37.4060

Expected values are what is actually observed

EXPCTD	COL1
ROW1	37.1523
ROW2	43.3261
ROW3	252.4958
ROW4	547.9020
ROW5	64.6349
ROW6	16.6936
ROW7	61.1350
ROW8	113.4433
ROW9	308.3708
ROW10	50.0531
ROW11	100.9904
ROW12	117.7725
ROW13	998.3425
ROW14	138.5726
ROW15	175.6960
ROW16	45.3779

Appendix I Program #7

ROW17	132.6864
ROW18	1319.6
ROW19	134.0881
ROW20	287.8310
ROW21	217.3951
ROW22	132.6864
ROW23	689.5739
ROW24	134.0881
ROW25	175.6804
ROW26	358.5980
ROW27	76.2834
ROW28	78.9320
ROW29	674.9224
ROW30	276.6978
ROW31	1935.7
ROW32	303.2831

Relative missing data

REL_MISS	COL1
ROW1	3072.0
ROW2	3048.0
ROW3	3880.1

Total number of frequencies

TOTAL	COL1
ROW1	35072.1

APPENDIX II

This is a listing of SAS program input and output of a program that performs Latent Class Analysis via Generalized Linear Models. Program input is analogous to that of program 7 in Appendix I, i.e., a Design Matrix "X", a "C" matrix, and a *Beta* vector which is a vector of initial values; in addition, the observed table, y, is input. The iterative re-weighted least squares algorithm used consists of 22 brief lines of code. The rest of the program relates to data input ("y", "X", "C", and *Beta*) and print statements.

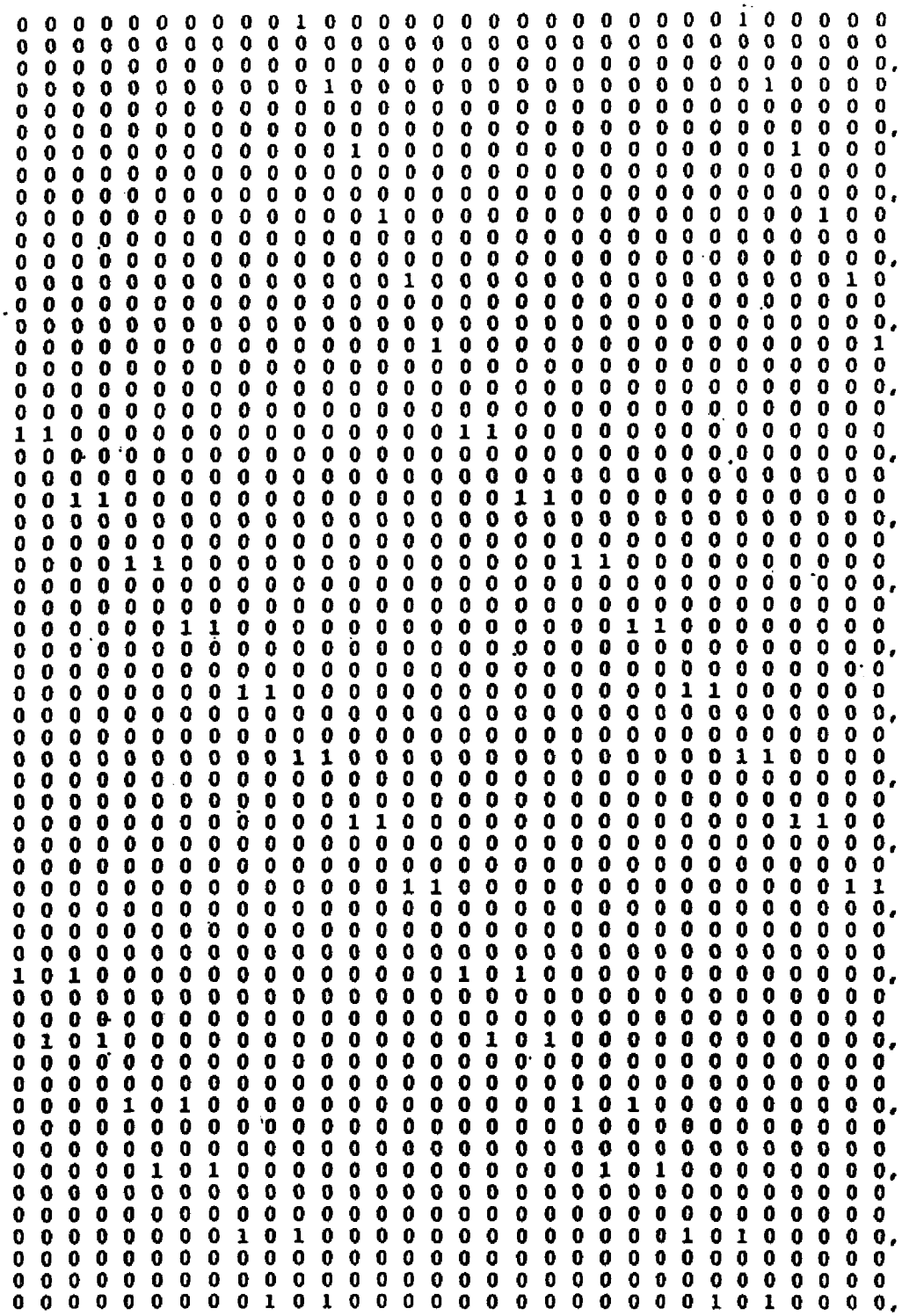
Appendix II

```
options lt linesize=80;
proc iml;


 /* CORRECT: MAR with unequal missing over B as in Little & Rubin */
*y = {129.6, 590.4, 214.3, 1815.8, 590.4, 129.6, 135.8, 214.2, 275.4, 174.6,
     100.8, 459.2, 2334.6, 275.4, 459.2, 100.8,
     80, 870, 80, 150, 50, 240, 290, 240};

*y = {144, 656,306,2594,656,144,194,306,306,194,144,656,2594,306,656,144};

 y = {144, 656,306,2594,656,144,194,306,306,194,144,656,2594,306,656,144,
       800, 2900, 800, 500, 500, 800, 2900, 800,
       450, 3250, 850, 450, 450, 850, 3250, 450};



 gen=  {   1   1   1   1   1,
           1  -1   1   1   1,
           1   1  -1   1   1,
           1  -1  -1   1   1,
           1   1   1  -1   1,
           1  -1   1  -1   1,
           1   1  -1  -1   1,
           1  -1  -1  -1   1,
           1   1   1   1  -1,
           1  -1   1   1  -1,
           1   1  -1   1  -1,
           1  -1  -1   1  -1,
           1   1   1  -1  -1,
           1  -1   1  -1  -1,
           1   1  -1  -1  -1,
           1  -1  -1  -1  -1,
          -1   1   1   1   1,
          -1  -1   1   1   1,
          -1   1  -1   1   1,
          -1  -1  -1   1   1,
          -1   1   1  -1   1,
          -1  -1   1  -1   1,
          -1   1  -1  -1   1,
          -1  -1  -1  -1   1,
          -1   1   1   1  -1,
          -1  -1   1   1  -1,
          -1   1  -1   1  -1,
          -1  -1  -1   1  -1,
          -1   1   1  -1  -1,
          -1  -1   1  -1  -1,
          -1   1  -1  -1  -1,
          -1  -1  -1  -1  -1 };

 misa=j(32,1,-1);
 misp=j(32,1,1);
 misn=j(32,1,0);
  mis=(misa//misp//misn)||(misa//misn//misp);

xx=j(96,1,1)||mis||repeat(gen,3,1);
```

Appendix II

```
xx=xx||j(96,6,0);


   xx(|,9|)=xx(|,4|)#xx(|,5|);
   xx(|,10|)=xx(|,4|)#xx(|,6|);
   xx(|,11|)=xx(|,4|)#xx(|,7|);
   xx(|,12|)=xx(|,4|)#xx(|,8|);


        /* MAR wrt D */
   xx(|,13|)=xx(|,2|)#xx(|,8|);
   xx(|,14|)=xx(|,3|)#xx(|,8|);


x=xx(|,{1 2 3 4 5 6 7 8 9 10 11 12 13 14 }|);

colms=
{"Int" "Miss A" "Miss B" "Class" "A" "B" "C" "D" "CL*A" "CL*B" "CL*C" "CL*D"
       "Miss A * D" "Miss B * D"};

print "Design Matrix" ,
   x(|colname=colms format=4.0|);


    /* Modelling Missing A and Missing B  Both MAR wrt D  */
 C = {1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
      0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
      0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
```

Appendix II



Appendix II

```
        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
        0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0,
        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
        0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1};

missing=(j(32,1,1)//j(64,1,0))||(j(32,1,0)//j(32,1,1)//j(32,1,0))||
         (j(64,1,0)//j(32,1,1));


* b1={2.1564,1.9924,-2.7E-16,4.1E-17,9.9E-18,-9.2E-17,-9.9E-17,0.6931,1.0986,
      -.693147,-1.0986, .67, .67 };* ,  .2 }; *,-.2,.3,-.3,.4,-.4,.5];

*b1={2.1557,1.6,1.6,0,0,0,0,0,1.0986,0.6931,
     -1.0986, -.693147, .67, .67 };

/* these were used to generate 1st pop */
*b1={5.0186,-.423649,-.761131,0,-.761131,0,.3375,.2027,-.337482,-.202733};


/* these were used to generate 2nd pop */

*b1={4.2767, 0.0, 0.0 ,0,.3,-.2,.1,-.3,0.6931,1.0986,
     -.693147, -1.0986,  .5, -.5 };

b1={3.2767, 0, 0, 0, 0 ,0, 0, 0, 0.6931, 1.0986,
     -.693147, -1.0986,  0, 0 };


x=xx(|,{1 2 3 4 5 6 7 8 9 10 11 12 13 14 }|);



start;
  do i=1 to 50;

  if i=1 then lr=.;
        lag_lr=lr;

    if i=1  then B=B1;
    eta = X*B;

     H1 = exp(eta);
      H = diag(h1);
     MU = C*H1;
      W = diag(1/MU);
      Z = (C*H*ETA)+(Y-MU);
           Xs = C*H*X;
           XpW = Xs'*W;
           T11 =XpW*Xs;
           T1 = ginv(t11);
           T2 = XpW*Z;
      B = T1*T2;

  LRM =  Y # log(Y/MU);
  LR  = 2 * LRM(|+,|);

dev_LR=lr-lag_lr;
```

Appendix II

```
  print "iteration =  " i "LR = " LR  "Deviation = " dev_lr ;

if abs(dev_lr) <.5  then goto endd;

   end;

endd:
finish;
run;


print "Convergence at Iteration= " i;
print "these are predicted freqs. in each cell of Underlying Table" ;
print hl;
print "Input Frequencies / Frequencies Predicted By the Model";
print y  mu;

print "Start Betas" bl "Final Betas" b;

 print lrm lr;

quit;
```

Appendix II

SAS 10:33 TUESDAY, OCTOBER 31, 1989 1

Design Matrix

X	Int	Miss A	Miss B	Clas s	A	B	C	D	CL*A	CL*B	CL*C	CL*D	Miss A * D	Miss B * D
ROW1	1	-1	-1	1	1	1	1	1	1	1	1	1	-1	-1
ROW2	1	-1	-1	1	-1	1	1	1	-1	1	1	1	-1	-1
ROW3	1	-1	-1	1	1	-1	1	1	1	-1	1	1	-1	-1
ROW4	1	-1	-1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
ROW5	1	-1	-1	1	1	1	-1	1	1	1	-1	1	-1	-1
ROW6	1	-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	-1
ROW7	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	-1
ROW8	1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	1	-1	-1
ROW9	1	-1	-1	1	1	1	1	-1	1	1	1	-1	1	1
ROW10	1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	1	1
ROW11	1	-1	-1	1	1	-1	1	-1	1	-1	1	-1	1	1
ROW12	1	-1	-1	1	-1	-1	1	-1	-1	-1	1	-1	1	1
ROW13	1	-1	-1	1	1	1	-1	-1	1	1	-1	-1	1	1
ROW14	1	-1	-1	1	-1	1	-1	-1	-1	1	-1	-1	1	1
ROW15	1	-1	-1	1	1	-1	-1	-1	1	-1	-1	-1	1	1
ROW16	1	-1	-1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1
ROW17	1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	-1	-1
ROW18	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	-1
ROW19	1	-1	-1	-1	1	-1	1	1	-1	1	-1	-1	-1	-1
ROW20	1	-1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1
ROW21	1	-1	-1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1
ROW22	1	-1	-1	-1	-1	1	-1	1	1	-1	1	-1	-1	-1
ROW23	1	-1	-1	-1	1	-1	-1	1	-1	1	1	-1	-1	-1
ROW24	1	-1	-1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1
ROW25	1	-1	-1	-1	1	1	1	-1	-1	-1	-1	1	1	1
ROW26	1	-1	-1	-1	-1	1	1	-1	1	-1	-1	1	1	1
ROW27	1	-1	-1	-1	1	-1	1	-1	-1	1	-1	1	1	1
ROW28	1	-1	-1	-1	-1	-1	1	-1	1	1	-1	1	1	1
ROW29	1	-1	-1	-1	1	1	-1	-1	-1	-1	1	1	1	1
ROW30	1	-1	-1	-1	-1	1	-1	-1	1	-1	1	1	1	1
ROW31	1	-1	-1	-1	1	-1	-1	-1	-1	1	1	1	1	1
ROW32	1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1
ROW33	1	1	0	1	1	1	1	1	1	1	1	1	1	0
ROW34	1	1	0	1	-1	1	1	1	-1	1	1	1	1	0
ROW35	1	1	0	1	1	-1	1	1	1	-1	1	1	1	0
ROW36	1	1	0	1	-1	-1	1	1	-1	-1	1	1	1	0
ROW37	1	1	0	1	1	1	-1	1	1	1	-1	1	1	0
ROW38	1	1	0	1	-1	1	-1	1	-1	1	-1	1	1	0
ROW39	1	1	0	1	1	-1	-1	1	1	-1	-1	1	1	0
ROW40	1	1	0	1	-1	-1	-1	1	-1	-1	-1	1	1	0
ROW41	1	1	0	1	1	1	1	-1	1	1	1	-1	-1	0
ROW42	1	1	0	1	-1	1	1	-1	-1	1	1	-1	-1	0
ROW43	1	1	0	1	1	-1	1	-1	1	-1	1	-1	-1	0
ROW44	1	1	0	1	-1	-1	1	-1	-1	-1	1	-1	-1	0
ROW45	1	1	0	1	1	1	-1	-1	1	1	-1	-1	-1	0
ROW46	1	1	0	1	-1	1	-1	-1	-1	1	-1	-1	-1	0
ROW47	1	1	0	1	1	-1	-1	-1	1	-1	-1	-1	-1	0
ROW48	1	1	0	1	-1	-1	-1	-1	-1	-1	-1	-1	-1	0
ROW49	1	1	0	-1	1	1	1	1	-1	-1	-1	-1	1	0
ROW50	1	1	0	-1	-1	1	1	1	1	-1	-1	-1	1	0
ROW51	1	1	0	-1	1	-1	1	1	-1	1	-1	-1	1	0
ROW52	1	1	0	-1	-1	-1	1	1	1	1	-1	-1	1	0
ROW53	1	1	0	-1	1	1	-1	1	-1	-1	1	-1	1	0

Appendix II

ROW54	1	1	0	-1	-1	1	-1	1	1	-1	1	-1	1	0
ROW55	1	1	0	-1	1	-1	-1	1	-1	1	1	-1	1	0
ROW56	1	1	0	-1	-1	-1	-1	1	1	1	1	-1	1	0
ROW57	1	1	0	-1	1	1	1	-1	-1	-1	-1	1	-1	0
ROW58	1	1	0	-1	-1	1	1	-1	1	-1	-1	1	-1	0
ROW59	1	1	0	-1	1	-1	1	-1	-1	1	-1	1	-1	0
ROW60	1	1	0	-1	-1	-1	1	-1	1	1	-1	1	-1	0
ROW61	1	1	0	-1	1	1	-1	-1	-1	-1	1	1	-1	0
ROW62	1	1	0	-1	-1	1	-1	-1	1	-1	1	1	-1	0
ROW63	1	1	0	-1	1	-1	-1	-1	-1	1	1	1	-1	0
ROW64	1	1	0	-1	-1	-1	-1	-1	1	1	1	1	-1	0
ROW65	1	0	1	1	1	1	1	1	1	1	1	1	0	1
ROW66	1	0	1	1	-1	1	1	1	-1	1	1	1	0	1
ROW67	1	0	1	1	1	-1	1	1	1	-1	1	1	0	1
ROW68	1	0	1	1	-1	-1	1	1	-1	-1	1	1	0	1
ROW69	1	0	1	1	1	1	-1	1	1	1	-1	1	0	1
ROW70	1	0	1	1	-1	1	-1	1	-1	1	-1	1	0	1
ROW71	1	0	1	1	1	-1	-1	1	1	-1	-1	1	0	1
ROW72	1	0	1	1	-1	-1	-1	1	-1	-1	-1	1	0	1
ROW73	1	0	1	1	1	1	1	-1	1	1	1	-1	0	-1
ROW74	1	0	1	1	-1	1	1	-1	-1	1	1	-1	0	-1
ROW75	1	0	1	1	1	-1	1	-1	1	-1	1	-1	0	-1
ROW76	1	0	1	1	-1	-1	1	-1	-1	-1	1	-1	0	-1
ROW77	1	0	1	1	1	1	-1	-1	1	1	-1	-1	0	-1
ROW78	1	0	1	1	-1	1	-1	-1	-1	1	-1	-1	0	-1
ROW79	1	0	1	1	1	-1	-1	-1	1	-1	-1	-1	0	-1
ROW80	1	0	1	1	-1	-1	-1	-1	-1	-1	-1	-1	0	-1
ROW81	1	0	1	-1	1	1	1	1	-1	-1	-1	-1	0	1
ROW82	1	0	1	-1	-1	1	1	1	1	-1	-1	-1	0	1
ROW83	1	0	1	-1	1	-1	1	1	-1	1	-1	-1	0	1
ROW84	1	0	1	-1	-1	-1	1	1	1	1	-1	-1	0	1
ROW85	1	0	1	-1	1	1	-1	1	-1	-1	1	-1	0	1
ROW86	1	0	1	-1	-1	1	-1	1	1	-1	1	-1	0	1
ROW87	1	0	1	-1	1	-1	-1	1	-1	1	1	-1	0	1
ROW88	1	0	1	-1	-1	-1	-1	1	1	1	1	-1	0	1
ROW89	1	0	1	-1	1	1	1	-1	-1	-1	-1	1	0	-1
ROW90	1	0	1	-1	-1	1	1	-1	1	-1	-1	1	0	-1
ROW91	1	0	1	-1	1	-1	1	-1	-1	1	-1	1	0	-1
ROW92	1	0	1	-1	-1	-1	1	-1	1	1	-1	1	0	-1
ROW93	1	0	1	-1	1	1	-1	-1	-1	-1	1	1	0	-1
ROW94	1	0	1	-1	-1	1	-1	-1	1	-1	1	1	0	-1
ROW95	1	0	1	-1	1	-1	-1	-1	-1	1	1	1	0	-1
ROW96	1	0	1	-1	-1	-1	-1	-1	1	1	1	1	0	-1

iteration =		LR =		Deviation =	
I	COL1	LR	COL1	DEV_LR	COL1
ROW1	1.0000	ROW1	64649.5	ROW1	64649.5

iteration =		LR =		Deviation =	
I	COL1	LR	COL1	DEV_LR	COL1
ROW1	2.0000	ROW1	-104956	ROW1	-169606

Appendix II

SAS 10:35 TUESDAY, OCTOBER 31, 1989 3

iteration =		LR =		Deviation =	
I	COL1	LR	COL1	DEV_LR	COL1
ROW1	3.0000	ROW1	-57333	ROW1	47623.2

iteration =		LR =		Deviation =	
I	COL1	LR	COL1	DEV_LR	COL1
ROW1	4.0000	ROW1	-24889	ROW1	32444.5

iteration =		LR =		Deviation =	
I	COL1	LR	COL1	DEV_LR	COL1
ROW1	5.0000	ROW1	-12354	ROW1	12534.5

iteration =		LR =		Deviation =	
I	COL1	LR	COL1	DEV_LR	COL1
ROW1	6.0000	ROW1	-18820	ROW1	-6466.1

iteration =		LR =		Deviation =	
I	COL1	LR	COL1	DEV_LR	COL1
ROW1	7.0000	ROW1	-8010.4	ROW1	10809.8

iteration =		LR =		Deviation =	
I	COL1	LR	COL1	DEV_LR	COL1
ROW1	8.0000	ROW1	-2098.5	ROW1	5911.9

iteration =		LR =		Deviation =	
I	COL1	LR	COL1	DEV_LR	COL1
ROW1	9.0000	ROW1	-174.55	ROW1	1924.0

iteration =		LR =		Deviation =	
I	COL1	LR	COL1	DEV_LR	COL1
ROW1	10.0000	ROW1	-1.7957	ROW1	172.7517

iteration =		LR =		Deviation =	
I	COL1	LR	COL1	DEV_LR	COL1
ROW1	11.0000	ROW1	-2.4E-04	ROW1	1.7955

iteration =		LR =		Deviation =	
I	COL1	LR	COL1	DEV_LR	COL1
ROW1	12.0000	ROW1	-2.5E-11	ROW1	2.4E-04

Appendix II

SAS 10:37 TUESDAY, OCTOBER 31, 1989 4

Convergence at Iteration=

I	COL1
ROW1	12.0000

these are predicted freqs. in each cell of Underlying Table

H1	COL1
ROW1	72.0000
ROW2	8.0000
ROW3	18.0000
ROW4	2.0000
ROW5	648.0000
ROW6	72.0000
ROW7	162.0000
ROW8	18.0000
ROW9	288.0000
ROW10	32.0000
ROW11	72.0000
ROW12	8.0000
ROW13	2592.0
ROW14	288.0000
ROW15	648.0000
ROW16	72.0000
ROW17	72.0000
ROW18	648.0000
ROW19	288.0000
ROW20	2592.0
ROW21	8.0000
ROW22	72.0000
ROW23	32.0000
ROW24	288.0000
ROW25	18.0000
ROW26	162.0000
ROW27	72.0000
ROW28	648.0000
ROW29	2.0000
ROW30	18.0000
ROW31	8.0000
ROW32	72.0000
ROW33	72.0000
ROW34	8.0000
ROW35	18.0000
ROW36	2.0000
ROW37	648.0000
ROW38	72.0000
ROW39	162.0000
ROW40	18.0000
ROW41	288.0000
ROW42	32.0000
ROW43	72.0000
ROW44	8.0000
ROW45	2592.0
ROW46	288.0000
ROW47	648.0000

Appendix II

ROW48	72.0000
ROW49	72.0000
ROW50	648.0000
ROW51	288.0000
ROW52	2592.0
ROW53	8.0000
ROW54	72.0000
ROW55	32.0000
ROW56	288.0000
ROW57	18.0000
ROW58	162.0000
ROW59	72.0000
ROW60	648.0000
ROW61	2.0000
ROW62	18.0000
ROW63	8.0000
ROW64	72.0000
ROW65	72.0000
ROW66	8.0000
ROW67	18.0000
ROW68	2.0000
ROW69	648.0000
ROW70	72.0000
ROW71	162.0000
ROW72	18.0000
ROW73	288.0000
ROW74	32.0000
ROW75	72.0000
ROW76	8.0000
ROW77	2592.0
ROW78	288.0000
ROW79	648.0000
ROW80	72.0000
ROW81	72.0000
ROW82	648.0000
ROW83	288.0000
ROW84	2592.0
ROW85	8.0000
ROW86	72.0000
ROW87	32.0000
ROW88	288.0000
ROW89	18.0000
ROW90	162.0000
ROW91	72.0000
ROW92	648.0000
ROW93	2.0000
ROW94	18.0000
ROW95	8.0000
ROW96	72.0000

Input Frequencies / Frequencies Predicted By the Model

Y	COL1	MU	COL1
ROW1	144.0000	ROW1	144.0000
ROW2	656.0000	ROW2	656.0000

Appendix II

ROW3	306.0000	ROW3	306.0000
ROW4	2594.0	ROW4	2594.0
ROW5	656.0000	ROW5	656.0000
ROW6	144.0000	ROW6	144.0000
ROW7	194.0000	ROW7	194.0000
ROW8	306.0000	ROW8	306.0000
ROW9	306.0000	ROW9	306.0000
ROW10	194.0000	ROW10	194.0000
ROW11	144.0000	ROW11	144.0000
ROW12	656.0000	ROW12	656.0000
ROW13	2594.0	ROW13	2594.0
ROW14	306.0000	ROW14	306.0000
ROW15	656.0000	ROW15	656.0000
ROW16	144.0000	ROW16	144.0000
ROW17	800.0000	ROW17	800.0000
ROW18	2900.0	ROW18	2900.0
ROW19	800.0000	ROW19	800.0000
ROW20	500.0000	ROW20	500.0000
ROW21	500.0000	ROW21	500.0000
ROW22	800.0000	ROW22	800.0000
ROW23	2900.0	ROW23	2900.0
ROW24	800.0000	ROW24	800.0000
ROW25	450.0000	ROW25	450.0000
ROW26	3250.0	ROW26	3250.0
ROW27	850.0000	ROW27	850.0000
ROW28	450.0000	ROW28	450.0000
ROW29	450.0000	ROW29	450.0000
ROW30	850.0000	ROW30	850.0000
ROW31	3250.0	ROW31	3250.0
ROW32	450.0000	ROW32	450.0000

Start Betas B1	COL1	Final Betas B	COL1
ROW1	3.2767	ROW1	4.2767
ROW2	0	ROW2	1.8E-16
ROW3	0	ROW3	9.6E-17
ROW4	0	ROW4	7.0E-16
ROW5	0	ROW5	-2.9E-16
ROW6	0	ROW6	-1.4E-16
ROW7	0	ROW7	3.0E-16
ROW8	0	ROW8	4.1E-17
ROW9	0.6931	ROW9	1.0986
ROW10	1.0986	ROW10	0.6931
ROW11	-.693147	ROW11	-1.0986
ROW12	-1.0986	ROW12	-.693147
ROW13	0	ROW13	-4.2E-17
ROW14	0	ROW14	4.6E-17

Appendix II

SAS 10:37 TUESDAY, OCTOBER 31, 1989 7

LRM	COL1	LR	COL1
ROW1	-3.6E-13	ROW1	-2.5E-11
ROW2	-1.4E-12		
ROW3	-1.3E-14		
ROW4	7.9E-13		
ROW5	-4.6E-13		
ROW6	-3.8E-13		
ROW7	-5.7E-13		
ROW8	-6.5E-13		
ROW9	-4.7E-13		
ROW10	-5.9E-13		
ROW11	-3.4E-13		
ROW12	-5.8E-13		
ROW13	1.4E-12		
ROW14	4.2E-14		
ROW15	-1.3E-12		
ROW16	-3.9E-13		
ROW17	-1.6E-12		
ROW18	1.4E-12		
ROW19	-6.4E-13		
ROW20	-1.1E-12		
ROW21	-9.4E-13		
ROW22	-7.5E-13		
ROW23	2.0E-12		
ROW24	-1.5E-12		
ROW25	-3.2E-13		
ROW26	-2.7E-13		
ROW27	-9.2E-13		
ROW28	-9.8E-13		
ROW29	-7.6E-13		
ROW30	-1.1E-12		
ROW31	4.5E-13		
ROW32	-3.0E-13		

APPENDIX III

This appendix contains a SAS IML program that performs maximum likelihood Latent Structure Analysis via the EM algorithm. It is documented internally. In terms of data input there are similarities to MLLSA (Clogg, 1977). Briefly:

1) A column frequency vector of the observed frequencies is read into the program; the order of data input is such that the leftmost subscript is varied the fastest (e.g., P_{1111}, P_{2111}, etc.). This vector is called *freq*. In the sample program the data is read in from an external file. The data could also be stacked with the program by modifying slightly the input statements.

2) Initial Latent Class probabilities, *pi_tX*, are input.

3) Initial conditional probabilities, *pi_cond*, are input.

4) A matrix, called *matrx*, tells the program which frequencies are to be summed to form the observed marginal frequencies.

5) The number of observed variables, *num_var*,
the number of levels of each observed variable, *num_lev*,
the number of latent classes, *num_class*,
the maximum number of iterations, *num_iter*,
the convergence criteria, *convrg*, are also given by the user.

6) If restrictions are required, they are programmed directly by the user.

```
filename in '[stewart.research.pops]mcar.dat';
options lt linesize=80;
data table;
  infile in;
  length dum1 $5. dum2 $6. dum3 $5.;
  drop dum1 dum2 dum3;
  input dum1  dum2 dum3   freq1  ;

  if  1 le _n_ le 32;

/* input observed frequency table as a column vector in the following order:

              P1111..., P2111..., P1211..., P2211...

       e.g., vary the first subscript the fastest and the last the slowest */
;
proc iml;
use table ;
read all;

print freq;

n=freq(|+,|);

print n;


*************************  INPUT PARAMETER VALUES HERE  **********************;

num_var=5;                 /* Number of Observed Variables */
num_lev={2 2 2 2 2};       /* Number of Levels For Each Observed Variable */
num_clas=4;                /* Number of Latent Classes */
num_iter=200;              /* Maximum Number of Iterations */
convrg=.001;               /* Convergence criteria */
prntstat='NO';             /* 'YES' Prints Intermediate Stats Every 20 Iter. */


  /* this matrix in transpose form indicates by '1's which frequencies
     are to be summed to form a given marginal frequency                 */


matrx ={
1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0, /* a1    */
0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1, /* a2    */
1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0, /* b1    */
0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1, /* b2    */
1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0, /* c1    */
0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1, /* c2    */
1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0, /* d1    */
0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1, /* d2    */
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0, /*comp. data*/
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1}; /*miss..data*/


pi_tX={.25 .25 .25 .25} ;          /* initial latent group probabilities */



pi_cond={
```

```
          .1  .9   1   1 ,   /* initial marginal conditional probs. */
          .9  .1   0   0 ,   /* each column is a latent class       */
          .2  .8  .2  .8 ,   /* 1st row is A1, 2nd row is A2        */
          .8  .2  .8  .2 ,   /* 3rd row is B1, etc.                 */
          .9  .1  .9  .1 ,
          .1  .9  .1  .9 ,
          .8  .2  .8  .2 ,
          .2  .8  .2  .8 ,
           1   1   0   0 ,
           0   0   1   1 );


****************************  END INPUT VALUES  ************************;

margfreq=matrx*freq;

p_obser=freq/N;

start iterat;

do iter = 1 to num_iter;

if iter=1 then do;

  Print "MAXIMUM   LIKELIHOOD   LATENT   STRUCTURE   ANALYSIS";
  Print "Version 1.1, October 1989 (William Stewart, author)";
  Print " ";
  Print "This is a "  num_class(|format=2.0|) " class, " num_var(|format=2.0|)
        " variable model.";
  Print " ";
  Print "Initial Observed Frequencies "     freq(|format=10.0|) ;
  Print "Marginal Observed Frequencies (NOTE: vars. are in inverse order)";
  Print       margfreq(|format=10.0|);
  Print "Total Sample Size" n(|format=11.0|);

  Print "Initial Conditional Probabilities " pi_cond;
  Print "Initial Latent Class Probabilities " pi_tX;
end;

  if iter=1 then lag_lr =0;
  else lag_lr=lr_chisq;


DO J= 1 to num_class;
DO I = 1 to num_var;
  if i=1 then m=0;
    l= m + 1 ;
    K = num_lev(|1,I|);
      m = m + k;
   if i=1 then pi_temp = pi_tx(|,J|)@pi_cond(|l:m,J|);
      else     pi_temp= pi_cond(|l:m,J|)@pi_temp;
  end;
  if j=1 then pi_x=pi_temp;
     else pi_x=pi_x||pi_temp;
 end;
```

Appendix III

```
pi_e_obs=pi_X(|,+|);                /* (2) estimated observed cell probs. */

pi_Xbar=(1/pi_e_obs)#pi_X ;         /* (3) conditional prob. of being in */
                                    /*  latent class given the observed cell*/


pi_ttemp=p_obser#pi_Xbar;

pi_tX = pi_ttemp(|+,|);             /* (4) latent class probabilities    */


pi_cond = (matrx*pi_ttemp)#(1/pi_tX); /* (5) new estimated conditional probs. */

lr=(freq#(1/N))#log((freq#(1/N))/pi_e_obs);
lr_chisq=2*N*lr(|+,|);

lr_dev=lag_lr-lr_chisq;

if iter>1 then do;
if mod(iter,10)=0 then do;
print "Iteration = " iter "    Decrease in Likelihood Ratio = " lr_dev;
end; end;

if mod(iter,20)=0 then do;
if  prntstat='YES' then do;


print "Intermediate Calculations For Iteration =       " iter;
print "Estimated Probabilities of Latent Cells         " pi_x;
print "Estimated Probabilities of Observed Cells       " pi_e_obs ;
print "Estimated Latent Class Probabilities            " pi_tX;
print "Estimate Conditional Probabilities              " pi_cond;
  end; end;

****************************** Begin Restrictions ************************;


    /* restrictions on conditional probabilities */

/*
pi_cond(|9,1|)=1;
pi_cond(|9,2|)=1;
pi_cond(|9,3|)=0;       examples of conditionals set to specific values
pi_cond(|9,4|)=0;
pi_cond(|9,5|)=0;
pi_cond(|9,6|)=0;
*/

   /* The following restrictions set conditional equal to one another  */

          /* Restrictions on B  */

pi_cond(|3,1|)=( pi_tx(|1,1|)*pi_cond(|3,1|)+pi_tx(|1,3|)*pi_cond(|3,3|) ) /
                ( pi_tx(|1,1|)+pi_tx(|1,3|));
pi_cond(|3,3|)=pi_cond(|3,1|);

pi_cond(|4,1|)=1-pi_cond(|3,1|);
pi_cond(|4,3|)=pi_cond(|4,1|);
```

Appendix III

```
pi_cond(|3,2|)=( pi_tx(|1,2|)*pi_cond(|3,2|)+pi_tx(|1,4|)*pi_cond(|3,4|) ) /
                ( pi_tx(|1,2|)+pi_tx(|1,4|));
pi_cond(|3,4|)=pi_cond(|3,2|);

pi_cond(|4,2|)=1-pi_cond(|3,2|);
pi_cond(|4,4|)=pi_cond(|4,2|);


          /* Restrictions on C  */

pi_cond(|5,1|)=( pi_tx(|1,1|)*pi_cond(|5,1|)+pi_tx(|1,3|)*pi_cond(|5,3|) ) /
                ( pi_tx(|1,1|)+pi_tx(|1,3|));
pi_cond(|5,3|)=pi_cond(|5,1|);

pi_cond(|6,1|)=1-pi_cond(|5,1|);
pi_cond(|6,3|)=pi_cond(|6,1|);

pi_cond(|5,2|)=( pi_tx(|1,2|)*pi_cond(|5,2|)+pi_tx(|1,4|)*pi_cond(|5,4|) ) /
                ( pi_tx(|1,2|)+pi_tx(|1,4|));
pi_cond(|5,4|)=pi_cond(|5,2|);

pi_cond(|6,2|)=1-pi_cond(|5,2|);
pi_cond(|6,4|)=pi_cond(|6,2|);

          /* Restrictions on D  */
pi_cond(|7,1|)=( pi_tx(|1,1|)*pi_cond(|7,1|)+pi_tx(|1,3|)*pi_cond(|7,3|) ) /
                ( pi_tx(|1,1|)+pi_tx(|1,3|));
pi_cond(|7,3|)=pi_cond(|7,1|);

pi_cond(|8,1|)=1-pi_cond(|7,1|);
pi_cond(|8,3|)=pi_cond(|8,1|);

pi_cond(|7,2|)=( pi_tx(|1,2|)*pi_cond(|7,2|)+pi_tx(|1,4|)*pi_cond(|7,4|) ) /
                ( pi_tx(|1,2|)+pi_tx(|1,4|));
pi_cond(|7,4|)=pi_cond(|7,2|);

pi_cond(|8,2|)=1-pi_cond(|7,2|);
pi_cond(|8,4|)=pi_cond(|8,2|);


*************************** End of Input Restrictions   **********************;


if abs(lr_dev)<convrg then goto fin;

end;

fin:
 finish;

run iterat;

iter=iter-1;

if abs(lr_dev)>convrg then

     print "No Convergence After Iterations =" Iter ;

if abs(lr_dev)<=convrg then
```

Appendix III

```
    print "Convergence After Iterations =" Iter ;

fin_freq=N#pi_e_obs;

sum1=freq(|+,|);
sum2=fin_freq(|+,|);
sum3=pi_e_obs(|+,|);

freqq=freq||fin_freq||pi_e_obs;

assign=pi_x(|,<:>|);
pi_x=pi_x||assign;

print "Estimated Final Latent Class Probabilities";
print pi_tX(|format=4.3|) ;
print "Estimated Final Latent Cell Probabilities/Latent Class Assignment";
print
"(WARNING: ties among classes will result in the cell placed in leftmost class)"
;
print pi_X(|format=6.5|) ;
print "Estimated Final Marginal Conditional (on Latent Class) Probabilities";
print pi_cond(|format=4.3|);
print
 "Actual Input Frequencies / Estimated Frequencies / Estimated Probabilities";
print freqq(| format=7.3|);
print sum1 sum2 sum3;

print "Final Likelihood Ratio Chi-square";
print lr_chisq(|format=7.2|);

stop;
quit;
```

Appendix III

SAS 14:28 TUESDAY, OCTOBER 31, 1989 1

FREQ	COL1
ROW1	0
ROW2	4.0000
ROW3	2.0000
ROW4	25.0000
ROW5	11.0000
ROW6	0
ROW7	2.0000
ROW8	7.0000
ROW9	6.0000
ROW10	2.0000
ROW11	0
ROW12	8.0000
ROW13	22.0000
ROW14	1.0000
ROW15	8.0000
ROW16	2.0000
ROW17	8.0000
ROW18	0
ROW19	29.0000
ROW20	0
ROW21	8.0000
ROW22	0
ROW23	8.0000
ROW24	0
ROW25	3.0000
ROW26	0
ROW27	3.0000
ROW28	0
ROW29	35.0000
ROW30	0
ROW31	6.0000
ROW32	0

N	COL1
ROW1	200.0000

MAXIMUM LIKELIHOOD LATENT STRUCTURE ANALYSIS

Version 1.1, October 1989 (William Stewart, author)

This is a class, variable model.

NUM_CLAS	CO L1	NUM_VAR	CO L1
ROW1	4	ROW1	5

Appendix III

SAS 14:29 TUESDAY, OCTOBER 31, 1989 2

Initial Observed Frequencies

FREQ	COL1
ROW1	0
ROW2	4
ROW3	2
ROW4	25
ROW5	11
ROW6	0
ROW7	2
ROW8	7
ROW9	6
ROW10	2
ROW11	0
ROW12	8
ROW13	22
ROW14	1
ROW15	8
ROW16	2
ROW17	8
ROW18	0
ROW19	29
ROW20	0
ROW21	8
ROW22	0
ROW23	8
ROW24	0
ROW25	3
ROW26	0
ROW27	3
ROW28	0
ROW29	35
ROW30	0
ROW31	6
ROW32	0

Marginal Observed Frequencies (NOTE: vars. are in inverse order)

MARGFREQ	COL1
ROW1	151
ROW2	49
ROW3	100
ROW4	100
ROW5	90
ROW6	110
ROW7	104
ROW8	96
ROW9	100
ROW10	100

Total Sample Size

Appendix III

SAS 14:29 TUESDAY, OCTOBER 31, 1989 3

N	COL1
ROW1	200

Initial Conditional Probabilities

PI_COND	COL1	COL2	COL3	COL4
ROW1	0.1000	0.9000	1.0000	1.0000
ROW2	0.9000	0.1000	0	0
ROW3	0.2000	0.8000	0.2000	0.8000
ROW4	0.8000	0.2000	0.8000	0.2000
ROW5	0.9000	0.1000	0.9000	0.1000
ROW6	0.1000	0.9000	0.1000	0.9000
ROW7	0.8000	0.2000	0.8000	0.2000
ROW8	0.2000	0.8000	0.2000	0.8000
ROW9	1.0000	1.0000	0	0
ROW10	0	0	1.0000	1.0000

Initial Latent Class Probabilities

PI_TX	COL1	COL2	COL3	COL4
ROW1	0.2500	0.2500	0.2500	0.2500

NOTE: ATTEMPTED LOG OF A ZERO OR NEGATIVE MATRIX ELEMENT.
RESULT SET TO ZERO.
NOTE: ATTEMPTED LOG OF A ZERO OR NEGATIVE MATRIX ELEMENT.
RESULT SET TO ZERO.
NOTE: ATTEMPTED LOG OF A ZERO OR NEGATIVE MATRIX ELEMENT.
RESULT SET TO ZERO.
NOTE: ATTEMPTED LOG OF A ZERO OR NEGATIVE MATRIX ELEMENT.
RESULT SET TO ZERO.
NOTE: ATTEMPTED LOG OF A ZERO OR NEGATIVE MATRIX ELEMENT.
RESULT SET TO ZERO.
NOTE: ATTEMPTED LOG OF A ZERO OR NEGATIVE MATRIX ELEMENT.
RESULT SET TO ZERO.
NOTE: ATTEMPTED LOG OF A ZERO OR NEGATIVE MATRIX ELEMENT.
RESULT SET TO ZERO.
NOTE: ATTEMPTED LOG OF A ZERO OR NEGATIVE MATRIX ELEMENT.
RESULT SET TO ZERO.

Convergence After Iterations =

ITER	COL1
ROW1	8.0000

Estimated Final Latent Class Probabilities

PI_TX	COL1	COL2	COL3	COL4
ROW1	.250	.250	.234	.266

Appendix III

SAS 14:29 TUESDAY, OCTOBER 31, 1989 4

Estimated Final Latent Cell Probabilities/Latent Class Assignment

(WARNING: ties among classes will result in the cell placed in leftmost class)

PI_X	COL1	COL2	COL3	COL4	COL5
ROW1	.00115	.00487	.00000	.00000	2
ROW2	.02425	.00012	.00000	.00000	1
ROW3	.00651	.00101	.00000	.00000	1
ROW4	.13788	.00003	.00000	.00000	1
ROW5	.00025	.04167	.00000	.00000	2
ROW6	.00537	.00105	.00000	.00000	1
ROW7	.00144	.00864	.00000	.00000	2
ROW8	.03053	.00022	.00000	.00000	1
ROW9	.00024	.01624	.00000	.00000	2
ROW10	.00500	.00041	.00000	.00000	1
ROW11	.00134	.00337	.00000	.00000	2
ROW12	.02844	.00008	.00000	.00000	1
ROW13	.00005	.13909	.00000	.00000	2
ROW14	.00111	.00350	.00000	.00000	2
ROW15	.00030	.02884	.00000	.00000	2
ROW16	.00630	.00073	.00000	.00000	1
ROW17	.00000	.00000	.02371	.00532	3
ROW18	.00000	.00000	.00000	.00000	1
ROW19	.00000	.00000	.13480	.00110	3
ROW20	.00000	.00000	.00000	.00000	1
ROW21	.00000	.00000	.00525	.04556	4
ROW22	.00000	.00000	.00000	.00000	1
ROW23	.00000	.00000	.02984	.00944	3
ROW24	.00000	.00000	.00000	.00000	1
ROW25	.00000	.00000	.00489	.01776	4
ROW26	.00000	.00000	.00000	.00000	1
ROW27	.00000	.00000	.02780	.00368	3
ROW28	.00000	.00000	.00000	.00000	1
ROW29	.00000	.00000	.00108	.15207	4
ROW30	.00000	.00000	.00000	.00000	1
ROW31	.00000	.00000	.00616	.03153	4
ROW32	.00000	.00000	.00000	.00000	1

Estimated Final Marginal Conditional (on Latent Class) Probabilities

PI_COND	COL1	COL2	COL3	COL4
ROW1	.045	.976	****	1
ROW2	.955	.024	.000	.000
ROW3	.150	.828	.150	.828
ROW4	.850	.172	.850	.172
ROW5	.819	.104	.819	.104
ROW6	.181	.896	.181	.896
ROW7	.829	.230	.829	.230
ROW8	.171	.770	.171	.770
ROW9	1	1	.000	.000
ROW10	.000	.000	****	1

Appendix III

SAS 14:29 TUESDAY, OCTOBER 31, 1989 5

Actual Input Frequencies / Estimated Frequencies / Estimated Probabilities

FREQQ	COL1	COL2	COL3
ROW1	0.000	1.202	0.006
ROW2	4.000	4.875	0.024
ROW3	2.000	1.505	0.008
ROW4	25.000	27.581	0.138
ROW5	11.000	8.384	0.042
ROW6	0.000	1.283	0.006
ROW7	2.000	2.016	0.010
ROW8	7.000	6.149	0.031
ROW9	6.000	3.296	0.016
ROW10	2.000	1.082	0.005
ROW11	0.000	0.942	0.005
ROW12	8.000	5.705	0.029
ROW13	22.000	27.828	0.139
ROW14	1.000	0.921	0.005
ROW15	8.000	5.827	0.029
ROW16	2.000	1.404	0.007
ROW17	8.000	5.806	0.029
ROW18	0.000	0.000	0.000
ROW19	29.000	27.180	0.136
ROW20	0.000	0.000	0.000
ROW21	8.000	10.161	0.051
ROW22	0.000	0.000	0.000
ROW23	8.000	7.858	0.039
ROW24	0.000	0.000	0.000
ROW25	3.000	4.530	0.023
ROW26	0.000	0.000	0.000
ROW27	3.000	6.297	0.031
ROW28	0.000	0.000	0.000
ROW29	35.000	30.631	0.153
ROW30	0.000	0.000	0.000
ROW31	6.000	7.537	0.038
ROW32	0.000	0.000	0.000

SUM1	COL1	SUM2	COL1	SUM3	COL1
ROW1	200.0000	ROW1	200.0000	ROW1	1.0000

Final Likelihood Ratio Chi-square

LR_CHISQ	COL1
ROW1	18.79

BIBLIOGRAPHY

Bergen, J. R. (1982). Latent class models in education. In Review of research in education, 10, edited by C. Clogg, 305-360.

Bishop, Y. M., Feinberg,S. E., & Holland P. W. (1975). Discrete multivariate analysis, Cambridge: MIT Press.

Blumenthal, S. (1968). Multinomial sampling with partially categorized data. Journal of the American Statistical Association, 63, 542-551.

Bock, R. D., & LIberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.

Chen, T., & Feinberg, S. E. (1976). The analysis of contingency tables with incompletely classified data. Biometrics 32 , 133-144.

Clogg, C. C. (1977). Unrestricted and restricted maximum likelihood latent structure analysis: A manual for users. University Park, PA: Population Issues Research Office Working Paper No. 1977-09.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Stat. Soc., Series B, 39 , 1-38.

Dixon, W. J. (1985). BMDP statistical software. Berkeley: Univ. Cal. Press.

Everitt, B. S. (1984). An introduction to latent variable models. London: Chapman and Hall.

Everitt, B. S. (1987). Introduction to optimization methods and their application in statistics. London: Chapman and Hall.

Everitt, B. S. & Hand , D. J. (1981). Finite mixture distributions. London: Chapman and Hall.

Fisher, R. A. (1930). Statistical methods for research workers, 3rd edition. Edinburgh: Oliver and Boyd.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. Journal of the American Statistical Association, 77, 270-278.

Goodmam, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentiafiable models. Biometrika, 61, 215-231.

Goodman, L. A. (1978). Analyzing qualitative/categorical data. Cambridge Mass.: Abt Books.

Grizzle, J. E., Starmer, C. F., and Koch, G. G. (1969). Analysis of categorical data by linear models. Biometrics, 25, 489-504.

Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data. Biometrics, 14, 174-194.

Hocking, R. R.., & Oxspring, H. H. (1974). The analysis of partially categorized contingency data. Biometrics, 30, 469-483.

Koch, G. G., Imrey, P. B., & Reinfurt, D. W. (1972). Linear model analysis of categorical data with incomplete response vectors. Biometrics, 28, 663-692.

Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: Wiley.

McCullagh, P., & Nelder, J. A. (1983). Generalized linear models. London: Chapman and Hall.

Muthén, B. (1979). Contributions to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. Journal of the Royal Statistical Society, Series A, 135, 308-313.

Nordheim, E. V. (1984). Inference from nonrandomly missing categorical data: An example from a genetic study on Turner's syndrome. Journal of the American Statistical Association, 79, 772-780.

Ploog, D. W. (1967). The behavior of squirrel monkeys as revealed by sociometry, bioacoustics, and brain stimulation. In Social comminications among primates, edited by S. Altmann, 149-184. Chicago: U. Chicago Press.

Rindskopf, D. (1983). Class notes for Qualitative Data Analysis. CUNY.

Rindskopf, D. (1984). Latent variable models: applications in education. In Contemporary educational psychology, 9. Edited by . 104-121.

Rindskopf, D. (1989, to be submitted for publication). A general approach to categorical data analysis with missing data, using generalized linear models with composite links.

SAS/IML user's guide (1985). Cary N.C. : SAS Institute.

Thompson, R., & Baker, R. J. (1981). Composite link functions in generalized linear models. Applied Statistics, 30, 125-131.

Woolson, R. F., & Clark, W. R. (1984). Analysis of categorical incomplete longitudinal data. Journal of the Royal Statistical Society, Series A, 147, Part 1, 87-99.