INFORMATION TO USERS

This manuscript has been reproduced 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.

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.

**ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600**

UMI[®]

ANALYSIS OF 2×2 TABLES OF COUNTS WITH BOTH COMPLETELY AND PARTIALLY CROSS-CLASSIFIED DATA

By
Isabelle Gaboury, B.Sc.
September 2001

A Thesis
submitted to the School of Graduate Studies and Research
in partial fulfillment of the requirements
for the degree of
Master of Science in Mathematics¹

© Copyright 2001
by Isabelle Gaboury, B.Sc., Ottawa, Canada

¹The M.Sc. Program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

**National Library
of Canada**

**Acquisitions and
Bibliographic Services**

**395 Wellington Street
Ottawa ON K1A 0N4
Canada**

**Bibliothèque nationale
du Canada**

**Acquisitions et
services bibliographiques**

**395, rue Wellington
Ottawa ON K1A 0N4
Canada**

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-66903-3

Canada

The undersigned hereby recommend to
the Faculty of Graduate Studies and Research
acceptance of the thesis

**ANALYSIS OF 2x2 TABLES OF COUNTS WITH BOTH
COMPLETELY AND PARTIALLY CROSS-CLASSIFIED DATA**

Submitted by
Isabelle Gaboury, B.Sc.

in partial fulfilment of the requirements
for the degree of Master of Science

Director, School of Mathematics and Statistics

Thesis Supervisor

External Examiner

Carleton University

September 2001

© 2001, Isabelle Gaboury

Abstract

In this thesis, we provide an exposition of various way of modeling incomplete categorical data. Given the pattern of missing data, we want to find the best model that fits a 2×2 table of counts with 2 supplemental margins. First, we show how the so-called expected information matrix is biased, even under a MAR mechanism of missing data. Secondly, we review five different families of models, grouped according to their assumptions about the missing data pattern. The five families allow us to cover the two ignorable missing data mechanisms: MCAR and MAR. We demonstrate how some families may be generalized to a 3×3 table of counts with supplemental margins. We implement four of the five families and present our results using examples found in the literature. We finally try to fit the best model to some data from the TRICC study.

Acknowledgements

I want to thank my supervisor, Dr. J.N.K. Rao for his availability, his help and useful discussions. Thank you for all you taught me. You have been a source of inspiration not only for this thesis, but for all my life. Thank you to Dr. George Wells for his help with epidemiological issues of this thesis and discussions about the data set. I would also like to thank Dr. Stuart G. Baker for his help and discussion in some of the computer programming.

Je voudrais aussi remercier mes parents pour leur soutien moral tout au long de mes études. Je sais que cette première séparation fut parfois difficile à vivre, mais elle m'a aidée à trouver ma voie. Merci à Philippe, pour ta patience et pour tous ces soirs d'encouragement quand Ottawa semblait trop loin.

Je voudrais aussi remercier le Fonds pour la Fondation de Chercheurs et l'Aide à la Recherche (FCAR). Sans leur soutien financier, cette thèse n'aurait pu voir le jour.

Merci à mes amis et à tous ceux et celles qui m'ont encouragé à persévérer; plus particulièrement à Julie, Sonia, Caroline, Julie, Sébastien, Geneviève et Marielle. Thanks to my friends in Ottawa: Gosia, Zahra and Lynn: moments spent with you are really precious! Finalement, merci à Dr. Bruno Rémillard pour m'avoir donné le goût de suivre ses traces et de vouloir toujours me surpasser.

Dedication

To my grand-father, Maurice

Contents

Abstract	iii
Acknowledgements	iv
Dedication	v
1 Introduction	1
1.1 Objectives	2
1.2 Outline of the thesis	2
2 Review	4
2.1 The complete counts	4
2.1.1 Measures of association	4
2.2 Missing data mechanism	6
2.2.1 Imputation-based and model-based procedures	6
2.3 The EM algorithm	8
2.4 Iterative proportional fitting algorithm (IPF)	14
2.5 Newton-Raphson's method (NR)	15
3 Modeling the missing data mechanism	18
3.1 MAR Process	18
3.1.1 Illustration	21
3.2 Family I (MCAR process)	25
3.2.1 The model	25

3.2.2	Iterative process	28
3.2.3	Multinomial sampling scheme related to Poisson distribution .	29
3.2.4	Variance formulae	31
3.2.5	Goodness-of-fit tests	32
3.3	Family II	33
3.3.1	The model	34
3.3.2	Goodness-of-fit tests	36
3.4	Family III	37
3.4.1	Goodman's model	37
3.4.2	Fay's model	42
3.4.3	Goodness-of-fit test	43
3.5	Family IV (Composite linear model)	44
3.5.1	The model	44
3.5.2	Illustration	47
3.5.3	Maximization procedure	51
3.5.4	CLM and 2×2 tables of counts	54
3.5.5	Goodness-of-fit test	56
3.6	Family V (Closed-form estimates)	56
3.6.1	The model	56
3.6.2	Goodness-of-fit test	58
3.6.3	Estimates	59
3.6.4	Boundary estimates	64
4	Examples	67
4.1	Family I (Chen and Fienberg model)	67
4.2	Family II (Stasny's model)	68
4.3	Family III (Fay's model)	70
4.4	Family IV (Baker's model)	74
4.5	Family V (Baker's model)	74
5	Application to the TRICC study	76
5.1	The TRICC study	76

5.1.1	Data	77
5.1.2	Results	78
5.1.3	Analysis	79
6	Conclusion	82
A	S-Plus Programs	84
A.1	Chen and Fienberg model	84
A.2	Stasny's model	85
A.3	Fay's model	94
A.4	Baker's model (Family V)	96
B	Parameter estimation under Stasny's model	104
	Bibliography	110

List of Tables

3.1	Underlying probabilities for a 2×2 table with two sets of partially classified margins	26
3.2	Observed data under Stasny's model	34
3.3	Expected cell counts for observed data under Stasny's model	37
3.4	Notation for $n_1 = n_2 = 2$	57
3.5	Parameterization according to model 3.54	58
3.6	Meaning of the model and MSS's under the complete data	60
3.7	Closed-form models	62
3.8	Boundary solutions	65
4.1	Two-way contingency table with two supplemental margin	68
4.2	Parameter estimates (standard errors) and expected data counts for fitted model to the data in Table 4.1	68
4.3	Observed LFS data for August to September 1979	69
4.4	Estimates of cell counts, π 's, and transition probabilities p_{ij} 's for August to September 1979	69
4.5	Parameter estimates and G^2 statistics for August to September 1979 .	70
4.6	Data on mental and physical status	70
4.7	Birth weight and smoking: observed counts	74
4.8	Estimated expected counts under Model $(\alpha_j, \beta_{..})$	75
5.1	Original allocation of patients for the TRICC study	78
5.2	Allocation of patients after randomization for the TRICC study . . .	78
5.3	Parameters (naive standard errors, unconditional standard errors) . .	79

5.4	Model fit (complete data counts; OR (s.e.), odds ratio (standard error); G^2 , likelihood ratio test statistic; d.f., degrees of freedom; p, p-value)	80
-----	---	----

List of Figures

3.1	Diagram of the causal system among variables A, B, C, and D	38
3.2	Causal diagram for equations (3.33) and (3.35)	41
3.3	Diagram of the causal system among variables A, B, and C	42
4.1	Example of an estimable model of nonresponse : relationship between physical and mental status	71

Chapter 1

Introduction

Nonresponse, which poses difficult questions of inference, is usually the consequence of human behavior. Most frequently, nonresponse in observational studies reflects unwillingness or reluctance of respondents to provide the requested information. However, many other factors may also contribute to nonresponse. These include the inability to understand a question, the mistaken belief that a question does not apply to them, or the unavailability of the information either from lapse of memory or lack of knowledge. Nonresponse may propagate itself: failure to respond to one question may induce further nonresponse to subsequent items.

Causes of nonresponse are not restricted solely to the behavior of respondents but may also arise from other aspects of the study execution. For example, interviewers may inadvertently skip questions through misunderstanding of the questionnaire or avoid questions that embarrass them. Longitudinal studies that collect information over time may have information missing for specific periods because of inability to reach respondents on one or more occasions. Briefly, the circumstances that lead to nonresponse are potentially varied and complex (Fay, 1986).

We classify missing data into three categories: (1) structural, when investigators face a latent (unobserved) variable, parameters are identifiable because of a structure imposed on the underlying distribution; (2) planned, when it is hard or expensive to make observations on a variable, such as censored data or information obtained from a surrogate variable; and (3) unplanned, when one or more variables are unintentionally

missing, such as nonresponse in surveys or clinical trials.

1.1 Objectives

It would seem appealing to discard incomplete observations and analyze only complete data. This method could be carried out and may give interesting results if it is applied in the situations where the amount of missing data is small. However, it could lead to important biases and it is in general not really efficient since we face a potential loss of information. For this reason, we have tried to make use of the partially observed data as well. For instance, odds ratio are generally lower when we simply use the completely observed data. By taking advantage of margins (partially observed information), one may avoid rejecting a hypothesis or declaring that a treatment is inefficient. This bias is more important especially if the pattern of missing data is missing at random (MAR). In this case, estimation of parameter and measures of association are definitively distorted. Then, it is important to develop models which take account of the missing data mechanism and consequently reflect the implication of nonresponse in the data.

The scope of this thesis is limited to categorical data. Inspired by the paper of Molenberghs (1999a), we review literature concerning models which fit missing data. Our goal is to find a suitable model that fits the trial of Transfusion Requirements In Critical Care (TRICC) study data well. The TRICC study is a multicenter, randomized controlled Canadian trial on red-cell transfusion. To do so, we implement algorithms on five different families of modeling missing data. Families are categorized based on their own assumptions on missing data mechanism.

1.2 Outline of the thesis

The outline of the thesis is as follows. Chapter 2 provides the technique on how statistical analysis is done with a 2×2 complete table of counts. It covers different measures of association well known in epidemiology, such as the odds ratio and relative risk. We also explain various pattern of missing data and give a brief summary of

the simplest way of dealing with missing data: imputation. We finally display three main algorithms which are the key algorithms used in the thesis. We show how those three algorithms perform with tables of complete counts. Chapter 3 is the core of the thesis. The reader will find different ways of modeling the missing data. We group them into five important families. Chapter 4 reproduces examples shown in the literature to illustrate how we implemented the algorithms. Chapter 5 analyses the TRICC study through the five families. Finally, Chapter 6 of the thesis discusses some open problems.

Chapter 2

Review

2.1 The complete counts

In this chapter, we consider a 2×2 table of counts $\{n_{ij}; i = 1, 2; j = 1, 2\}$, with row margins $\{n_{i+}\}$ fixed. The row margins correspond to factor $i = 1, 2$ while the columns j refer to response groups (e.g. absence or presence of side effects). Such tables arise in comparative studies.

2.1.1 Measures of association

In comparative studies, it is important to study the impact of given factors on different response groups of individuals. Measures of association are utilized for this purpose. One measure, called the relative risk (RR), is defined as the ratio of the “success” probabilities for the two different groups. The RR is represented as $\pi_{1|1}/\pi_{1|2}$ where $\pi_{1|i}$ is the conditional probability of success of group 1 given factor i ($= 1, 2$). If the null hypothesis of no association H_0 is true, i.e. $\pi_{1|1} = \pi_{1|2}$, then $\pi_{1|1}/\pi_{1|2} = 1$ and we conclude that the response is not affected by the factor. In order to estimate the RR, one simply replaces $\pi_{1|1}/\pi_{1|2}$ by the sample RR, say $p_{1|1}/p_{1|2}$, where $p_{1|i} = n_{1i}/n_{i+}$ is the proportion of “successes” in a sample of size n_{i+} at factor i .

We can also derive a $(1-\alpha)$ -level confidence interval for RR using $\log(p_{1|1}/p_{1|2})$, because the distribution of $p_{1|1}/p_{1|2}$ could be highly skewed unless the sample sizes are quite

large. In this formula, $\log(p_{1|1}/p_{1|2})$ is approximately normal with mean $\log(\pi_{1|1}/\pi_{1|2})$ and variance $\sqrt{\frac{1-p_{1|1}}{n_{1+}p_{1|1}} + \frac{1-p_{1|2}}{n_{2+}p_{1|2}}}$. A $(1-\alpha)$ -level confidence interval on $\log(\pi_{1|1}/\pi_{1|2})$ is therefore given by

$$\log(p_{1|1}/p_{1|2}) \pm Z_{\alpha/2} \sqrt{\frac{1-p_{1|1}}{n_{1+}p_{1|1}} + \frac{1-p_{1|2}}{n_{2+}p_{1|2}}}, \quad (2.1)$$

where $Z_{\alpha/2}$ is the upper $\alpha/2$ -point of $N(0, 1)$ variable. Because of invariance, we can obtain a $(1-\alpha)$ -level confidence interval for $RR = (p_{1|1}/p_{1|2})$ by exponentiating the interval found in (2.1).

A second measure of association for 2×2 tables is the odds ratio (OR) defined as:

$$\theta = \frac{\pi_{1|1}/\pi_{2|1}}{\pi_{1|2}/\pi_{2|2}} = \frac{\pi_{1|1}\pi_{2|2}}{\pi_{1|2}\pi_{2|1}}.$$

Obviously, we can estimate θ by utilizing the sample OR:

$$\hat{\theta} = \frac{p_{1|1}p_{2|2}}{p_{1|2}p_{2|1}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Here again, it is feasible to get a $(1-\alpha)$ -level confidence interval for the OR by considering $\log(\theta)$ for the same reasons mentioned above for the RR. In this case, $\log(\hat{\theta})$ is approximately normal with mean $\log(\theta)$ and variance

$$\sqrt{\frac{1}{n_{1+}p_{1|1}} + \frac{1}{n_{1+}(1-p_{1|1})} + \frac{1}{n_{2+}p_{1|2}} + \frac{1}{n_{2+}(1-p_{1|2})}}. \quad (2.2)$$

Using the delta method, a $(1-\alpha)$ -level confidence interval on θ is given by

$$\begin{aligned} & \log(\hat{\theta}) \pm Z_{\alpha/2} \sqrt{\frac{1}{n_{1+}p_{1|1}} + \frac{1}{n_{1+}(1-p_{1|1})} + \frac{1}{n_{2+}p_{1|2}} + \frac{1}{n_{2+}(1-p_{1|2})}} \\ &= \log(\hat{\theta}) \pm Z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \end{aligned} \quad (2.3)$$

As for the RR, we have to exponentiate (2.3) in order to get a $(1-\alpha)$ -level confidence interval on θ .

It is also important to estimate the variance of $\hat{\theta} = \exp(\hat{\tau})$. Using the delta method, we have

$$\text{var}(\hat{\theta}) = \hat{\theta}^2 \text{var}(\hat{\tau}), \quad (2.4)$$

where $\text{var}(\hat{\tau})$ is given by (2.2).

Note finally that RR and OR can be linked in the case of rare diseases. In fact,

$$OR = \frac{\pi_{1|1}\pi_{2|2}}{\pi_{1|2}\pi_{2|1}} = \frac{\pi_{1|1}(1 - \pi_{1|2})}{\pi_{1|2}(1 - \pi_{1|1})} = RR \left(\frac{1 - \pi_{1|2}}{1 - \pi_{1|1}} \right)$$

and both $\pi_{1|1}$ and $\pi_{1|2}$ are close to zero under rare disease assumption. Therefore, we may say $OR \approx RR$, and this states the relationship between OR and RR.

2.2 Missing data mechanism

An important factor in choosing an appropriate method to analyze an incomplete data set is the mechanism which leads to missing values. The missing data mechanism should be well assessed since any analysis of the data depends on it. We classify missing data into two groups: ignorable and non-ignorable. Ignorable mechanisms can be split into two categories: (1) missing at random (MAR), and (2) missing completely at random (MCAR). Let X and Y be the variables of a bivariate dataset. The missing data mechanism on Y is said to be MAR if the probability of response on Y depends on X but not on Y . On the contrary, if the probability of response on Y is independent of X and Y , we will analyze the missing data as MCAR. This is equivalent to saying that missing data are MAR and observed data are observed at random (OAR). However, it is more common to use the term MCAR. The following section introduces different ways of dealing with missing data. We say that the mechanism is non-ignorable if it does not belong to any of the above categories.

2.2.1 Imputation-based and model-based procedures

An imputation-based procedure is used when missing data are filled in and then analyzed by standard methods. There are three main categories of imputation: (1)

hot and cold deck procedures, (2) mean imputation, and (3) regression imputation. In hot deck imputation, we substitute each missing value by a selected imputed value which belongs to the completely observed data set. We use the term cold deck when the imputed value comes from an external source such as values obtained from a previous survey on the same topic. Mean imputation is a kind of extension of hot deck imputation, since missing values are replaced by the means from sets of recorded values in the sample. Finally, with regression imputation, we estimate missing values by a prediction equation computed using known variables related to the former. For a best use of imputation routine, modification of standard analyses are sometimes required in order to allow for the differing status of real and imputed values.

The main advantages of imputation techniques are the following: (1) it is easier to specify the data structure using the terminology of experimental design, (2) it is easier to compute necessary statistical summaries, (3) it is easier to interpret the results of analyses since standard displays and summaries can be used, and (4) these methods are generally very flexible and straightforward to implement (Little and Rubin, 1987). However, these methods usually make strong assumptions such as the data are MCAR, and could lead to serious bias in some cases. In addition, and particularly for mean imputation, variance and covariance estimates are systematically underestimated since the actual sample size is reduced due to nonresponse. Even if we use adjustment factors, estimates of variance and covariance are not generally satisfactory. Regression imputation yields more reliable results than mean imputation, especially if normality assumption of the data set is plausible, but variance and covariance estimates are still underestimated. Moreover, this method has limitations when the missing variables are categorical because of the continuous nature of this procedure. In biostatistical research, imputation-based procedures tend to be less preferred for likelihood-based inference methods.

Statistical literature on missing data analysis abounds in procedures where a model is defined for partially missing data and where inferences are made from the likelihood under that model, with parameters estimated by procedures such as maximum likelihood (ML). There are many advantages in using model-based procedures. Some of them are: flexibility, the avoidance of ad hoc methods, assumptions underlying

the resulting methods can be displayed and evaluated, and the availability of large sample estimates of variances and covariances based on second derivatives of the log-likelihood, which take account of incompleteness in the data. The main focus of our work is on model-based procedures, which are presented in Chapter 3 as five different families.

2.3 The EM algorithm

The EM algorithm can handle a very broad range of problems concerning missing data since it was designed to answer intuitively the difficulties encountered by incomplete data sets. The basic idea used by this general iterative algorithm consists in: (1) replacing missing values by estimated values, (2) estimating the parameters, (3) reestimating the missing values assuming the new parameter estimates are correct, and finally (4) reestimating the parameters, and so forth, iterating until convergence. Let $f(\mathbf{Y}|\theta)$ be the density of a model for the complete data $\mathbf{Y}$, with θ being an unknown parameter (possibly vector-valued) estimated by $\hat{\theta}$. Then we can write $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ where $\mathbf{Y}_{obs}$ represents the observed part of $\mathbf{Y}$ and $\mathbf{Y}_{mis}$ denotes the missing values. In MAR process, we want to maximize the likelihood $L(\theta|\mathbf{Y}_{obs}) = \int f(\mathbf{Y}_{obs}, \mathbf{Y}_{mis}|\theta) d\mathbf{Y}_{mis}$ with respect to θ . Very often, a closed-form expression cannot be found from this equation. Therefore, one needs to use iterative methods and then maximize the log-likelihood $l(\theta^{(t)}|\mathbf{Y}_{obs})$, where $\theta^{(t)}$ is the estimate of θ at the t -th iteration. Consequently, these methods require an initial estimate of θ , denoted $\theta^{(0)}$. EM algorithm relates ML estimation of θ from $l(\theta|\mathbf{Y}_{obs})$ to ML estimation based on the complete-data log-likelihood $l(\theta|\mathbf{Y})$.

The main advantages of this algorithm are that these steps are generally easy to construct conceptually and have a direct statistical interpretation. In addition, it can be shown that EM algorithm converges reliably in the sense that under general conditions, each iteration increases the log-likelihood $l(\theta^{(t)}|\mathbf{Y}_{obs})$, and if $l(\theta|\mathbf{Y}_{obs})$ is bounded, the sequence $l(\theta^{(t)}|\mathbf{Y}_{obs})$ converges to a stationary value of $l(\theta|\mathbf{Y}_{obs})$. Generally, if the sequence of $\theta^{(t)}$ converges, it converges to a local maximum or saddle point of $l(\theta|\mathbf{Y}_{obs})$. An important disadvantage of EM algorithm is that the rate of

convergence could be very slow if the data set contains many missing observations. Dempster, Laird and Rubin (1977) showed that the convergence is linear with rate proportional to the fraction of information about θ in $l(\theta|\mathbf{Y})$, i.e.

$$|\theta^{(t+1)} - \hat{\theta}| = \lambda |\theta^{(t)} - \hat{\theta}|,$$

where λ is the ratio of the proportion of subjects from which we have missing observations to the proportion of subjects from which we have a complete set of observations. Each iteration of the EM algorithm is composed of an E-step (expectation step) and an M-step (maximization step). The expectation step finds the conditional expectation of a function of missing data $\mathbf{Y}_{mis}$ given the observed data $\mathbf{Y}_{obs}$ and the current estimate parameter, $\theta^{(t)}$, i.e. the E-step consists of computing

$$Q(\theta|\theta^{(t)}) = \int l(\theta|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta^{(t)}) d\mathbf{Y}_{mis}. \quad (2.5)$$

In the M-step, the Q function is maximized with respect to θ to obtain $\theta^{(t+1)}$. The algorithm is iterated until the desired convergence.

Theoretically, the log-likelihood of the complete data $\mathbf{Y}$ can be represented as

$$l(\theta|\mathbf{Y}) = l(\theta|\mathbf{Y}_{obs}, \mathbf{Y}_{mis}) = l(\theta|\mathbf{Y}_{obs}) + \ln f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta).$$

Since we want to estimate θ by maximizing the incomplete-data log-likelihood $l(\theta|\mathbf{Y}_{obs})$, we rewrite $l(\theta|\mathbf{Y}_{obs})$ as

$$l(\theta|\mathbf{Y}_{obs}) = l(\theta|\mathbf{Y}) - \ln f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta). \quad (2.6)$$

By taking the expectation of (2.6) on both sides with respect to $f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta^{(t)})$, we get

$$l(\theta|\mathbf{Y}_{obs}) = Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)}),$$

where

$$H(\theta|\theta^{(t)}) = \int \ln f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta) f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta^{(t)}) d\mathbf{Y}_{mis}.$$

and $Q(\theta|\theta^{(t)})$ as stated in (2.5).

Now, we need to state a well-known theorem (as explained by Silvey (1975)) in order to derive the next step.

Theorem 2.1 (Jensen's inequality) *If x is a random variable such that $E(x) = \mu$, and $f(x)$ is a convex function, then,*

$$E[f(x)] \geq f[E(x)]$$

with equality if and only if x is a degenerate distribution at μ .

Theorem 2.2 *Let q and r be density functions of two different probability distributions on the same probability space Y of points y , these distributions being different in the sense that there exists a set of positive q -probability on which $q(y) \neq r(y)$; and let C be any continuous convex function of a non-negative variable. Then,*

$$E_q \left[C \left(\frac{r}{q} \right) \right] \geq C(1),$$

with strictly inequality if C is strictly convex.

Proof. By Jensen's inequality, we have

$$E_q \left[C \left(\frac{r}{q} \right) \right] \geq C \left[E_q \left(\frac{r}{q} \right) \right]$$

and the inequality is strict if C is strictly convex, since r/q is not constant with q -probability 1 by assumption. The proof is now completed by notify that

$$\begin{aligned} E_q \left(\frac{r}{q} \right) &= \int_Y \frac{r(y)}{q(y)} q(y) dy, \\ &= \int_Y r(y) dy = 1, \end{aligned}$$

so that

$$E_q \left[C \left(\frac{r}{q} \right) \right] \geq C(1).$$

Note that the natural logarithm function is a continuous convex function of non-negative variables. Then, using the above theorem, we have:

$$E_{\theta^{(t)}} \left(-\ln \frac{f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta)}{f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta^{(t)})} \right) \geq -\ln(1) = 0$$

i.e.

$$H(\theta^{(t)}|\theta^{(t)}) - H(\theta|\theta^{(t)}) \geq 0,$$

or

$$H(\theta^{(t)}|\theta^{(t)}) \geq H(\theta|\theta^{(t)}). \quad (2.7)$$

Now, consider a sequence of iterates $\theta^{(0)}, \theta^{(1)}, \dots$, where $\theta^{(t+1)} = M(\theta^{(t)})$ for some function $M(\cdot)$. The difference in values of $l(\theta|\mathbf{Y}_{obs})$ at successive iterates is given by

$$\begin{aligned} l(\theta^{(t+1)}|\mathbf{Y}_{obs}) - l(\theta^{(t)}|\mathbf{Y}_{obs}) = \\ [Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})] - [H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})]. \end{aligned} \quad (2.8)$$

The ML estimator of $Q(\theta|\theta^{(t)})$ with respect to θ is $\theta^{(t+1)}$. More generally, the ML estimator of a generalized EM (GEM) is $\theta^{(t+1)}$. Then, $Q(\theta^{(t+1)}|\theta^{(t)}) > Q(\theta^{(t)}|\theta^{(t)})$. Hence, this implies that the first difference on the right hand side of (2.8) will always be greater than 0. In addition, according to (2.7), the second part of the right hand side of (2.8) is negative. Hence, we can conclude that $l(\theta^{(t+1)}|\mathbf{Y}_{obs}) - l(\theta^{(t)}|\mathbf{Y}_{obs}) > 0$ and so that the change from $\theta^{(t)}$ to $\theta^{(t+1)}$ increases the log-likelihood for EM or GEM algorithm. This completes the proof of the following theorem which is one of the important result of the EM algorithm stated by Little and Rubin (1987).

Theorem 2.3 *Every GEM algorithm increases $l(\theta|\mathbf{Y}_{obs})$ at each iteration, that is*

$$l(\theta^{(t+1)}|\mathbf{Y}_{obs}) \geq l(\theta^{(t)}|\mathbf{Y}_{obs}),$$

with equality if and only if

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}).$$

This theorem implies that the incomplete-data likelihood is nondecreasing on each iteration of a GEM algorithm, and is strictly increasing on any iteration such that Q increases (i.e., $Q(\theta^{(t+1)}|\theta^{(t)}, \mathbf{Y}_{obs}) > Q(\theta^{(t)}|\theta^{(t)}, \mathbf{Y}_{obs})$).

Corollary 1 Suppose that for some θ^* in the parameter space of θ , $l(\theta^*|\mathbf{Y}_{obs}) \geq l(\theta|\mathbf{Y}_{obs})$ for all θ . Then, for every GEM algorithm,

$$M(\theta^*) = \theta^*.$$

Corollary 2 Suppose that for some θ^* in the parameter space of θ , $l(\theta^*|\mathbf{Y}_{obs}) \geq l(\theta|\mathbf{Y}_{obs})$ for all θ , i.e. the convergence state is reached. Then, for every GEM algorithm,

$$\begin{aligned} l(M(\theta^*)|\mathbf{Y}_{obs}) &= l(\theta^*|\mathbf{Y}_{obs}) \\ Q(M(\theta^*)|\mathbf{Y}_{obs}) &= Q(\theta^*|\mathbf{Y}_{obs}) \end{aligned}$$

and

$$f(Y_{mis}|\mathbf{Y}_{obs}, M(\theta^*)) = f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \theta^*),$$

almost everywhere.

These corollaries imply that a maximum likelihood estimate (MLE) of θ is a fixed point of a GEM algorithm.

Theorem 2.4 Suppose a sequence of EM iterates are such that

(a) $D^{10}Q(\theta^{(t+1)}|\theta^{(t)}) = 0$, where D^{10} means the first derivative with respect to the first argument, that is

$$D^{10}Q(\theta^{(t+1)}|\theta^{(t)}) = \frac{\partial}{\partial\theta}Q(\theta|\theta^{(t)})|_{\theta=\theta^{(t+1)}},$$

(b) $\theta^{(t)}$ converges to θ^* , and

(c) $f(Y_{mis}|Y_{obs}, \theta)$ is smooth in θ , where smooth is defined in the proof. Then

$$Dl(\theta^*|Y_{obs}) = \frac{\partial}{\partial\theta}l(\theta|Y_{obs})|_{\theta=\theta^*} = 0,$$

so that if the $\theta^{(t)}$ converge, they converge to a stationary point.

Proof.

$$\begin{aligned} Dl(\theta^{(t+1)}|Y_{obs}) &= D^{10}Q(\theta^{(t+1)}|\theta^{(t)}) - D^{10}H(\theta^{(t+1)}|\theta^{(t)}) \\ &= -D^{10}H(\theta^{(t+1)}|\theta^{(t)}), \text{ by assumption} \\ &= \frac{\partial}{\partial\theta} \int \ln f(Y_{mis}|Y_{obs}, \theta) f(Y_{mis}|Y_{obs}, \theta^*) dY_{mis}, \end{aligned}$$

which, assuming sufficient smoothness to interchange the order of differentiation and integration, converges to

$$\int \frac{\partial}{\partial\theta} f(Y_{mis}|Y_{obs}, \theta^*) D,$$

which equals zero after interchanging the order of integration and differentiation.

Some other results concerning convergence of EM algorithm are formulated by Dempster, Laird and Rubin (1977) and Wu (1983). These results are:

1. If $l(\theta|Y_{obs})$ is bounded, $l(\theta^{(t)}|Y_{obs})$ converges to some θ^* .
2. If $f(Y|\theta)$ is a general (curved) exponential family and $l(\theta|Y_{obs})$ is bounded, then $l(\theta^{(t)}|Y_{obs})$ converges to a stationary value l^* .
3. If $f(Y|\theta)$ is a regular exponential family and $l(\theta|Y_{obs})$ is bounded, then $\theta^{(t)}$ converges to a stationary point θ^* .

2.4 Iterative proportional fitting algorithm (IPF)

The iterative proportional fitting (IPF) algorithm was first proposed by Deming and Stephan (1940). This method is in fact a kind of EM algorithm and it is used for calculating the estimates of expected counts, m_i , for loglinear models. As described by Agresti (1990), the procedure has the following steps:

1. Start with any initial estimates $\hat{m}_i^{(0)}$ of m_i having association and interaction structures no more complex than the model being fitted. For instance, $\hat{m}_i^{(0)} = 1$ satisfies all the models for two-way tables.
2. Successively adjust $\hat{m}_i^{(0)}$ by multiplying by appropriate scaling factors, so they match each marginal table in the set of minimal sufficient statistics (MSS).
3. Continue the process until the difference between the sufficient statistics and their fitted values is sufficiently closed to zero.

It is possible to illustrate the IPF algorithm by fitting the loglinear model to a three-way table of counts $\{n_{ijk}; i = 1, 2; j = 1, 2; k = 1, 2\}$ where the MSS's are the marginals n_{ij+} , n_{i+k} , and n_{+jk} , which correspond to no two- and three-factors interaction. Initial estimates must satisfy the model. The first cycle of the adjustment process has three steps:

$$\begin{aligned}\hat{m}_{ijk}^{(1)} &= \hat{m}_{ijk}^{(0)} \left[\frac{n_{ij+}}{\hat{m}_{ij+}^{(0)}} \right] \\ \hat{m}_{ijk}^{(2)} &= \hat{m}_{ijk}^{(1)} \left[\frac{n_{i+k}}{\hat{m}_{i+k}^{(1)}} \right] \\ \hat{m}_{ijk}^{(3)} &= \hat{m}_{ijk}^{(2)} \left[\frac{n_{+jk}}{\hat{m}_{+jk}^{(2)}} \right]\end{aligned}$$

Summing both sides of the first expression over k , we get $\hat{m}_{ij+}^{(1)} = n_{ij+}$. Thus, after this step, all observed and fitted frequencies match in the row counts. After the second step, all $\hat{m}_{i+k}^{(2)} = n_{i+k}$, but the fitted and observed row counts no longer

match. Finally, after the third step, we have $\hat{m}_{+jk}^{(2)} = n_{+jk}$, but again, the fitted and observed row counts no longer match. A new cycle begins by again matching the fitted and observed row and column counts using $\hat{m}_{ij}^{(4)} = \hat{m}_{ij}^{(3)} \left[\frac{n_{ij+}}{\hat{m}_{ij+}^{(3)}} \right]$, and so forth. As the cycles progress, the fitted values come simultaneously closer to matching the observed data in all sufficient statistics, and they exhibit those associations allowed in the model. We start a new cycle as long as we do not obtain the desired level of convergence.

2.5 Newton-Raphson's method (NR)

Newton-Raphson's (NR) method is a numerical procedure to find the maximum of a nonlinear function, such as the likelihood equation. As for the EM algorithm, the NR method requires an initial guess for the value that maximizes the function. Basically, the function is then approximated in a neighborhood of that guess by a second-degree polynomial and this second guess is now taken as the maximum value of the function. We repeat this procedure for a third guess and so forth. It follows that we obtain a sequence of guesses which converge to the location of the maximum provided that the function does not show too many irregularities and/or the first guess is good. Theoretically, we want to determine the maximum value $\hat{\phi}$ of ϕ of a general function $g(\phi)$, where ϕ is the parameter to estimate, possibly vector-valued. Let $q' = (\partial g / \partial \phi_1, \partial g / \partial \phi_2, \dots)$, and let H (which is also called the Hessian matrix) denotes the matrix having ab -th entries $h_{ab} = \partial^2 g / \partial \phi_a \partial \phi_b$. In the cases of incomplete data, both matrices H and q are conditional on the observed information, i.e. we can rewrite $h_{ab} = \partial^2 g(\phi | Y_{obs}) / \partial \phi_a \partial \phi_b$ and $q' = [\partial g(\phi | Y_{obs}) / \partial \phi_1, \partial g(\phi | Y_{obs}) / \partial \phi_2, \dots]$. Let $q^{(t)}$ and $H^{(t)}$ be the values of q and H evaluated at the t -th iteration. At the t -th iteration ($t = 0, 1, 2, \dots$), the function $g(\phi)$ is approximated near $\hat{\phi}^{(t)}$ by terms up to second order in its Taylor series expansion :

$$Q^{(t)}(\phi) = g(\phi^{(t)}) + q^{(t)'}(\phi - \phi^{(t)}) + \left(\frac{1}{2}\right) (\phi - \phi^{(t)})' H^{(t)} (\phi - \phi^{(t)}).$$

Solving $\partial Q^{(t)} / \partial \phi = q^{(t)} + H^{(t)}(\phi - \phi^{(t)}) = 0$ for ϕ yields the next iterate,

$$\phi^{(t+1)} = \phi^{(t)} - [H^{(t)}]^{-1} q^{(t)}, \quad (2.9)$$

assuming that $H^{(t)}$ is nonsingular.

For illustration, suppose that $g(\phi)$ is the log-likelihood for a Poisson model. This could be used to find the MLE in a complete table of counts. Let us represent that log-likelihood as

$$L(\phi) = \sum_i n_i \left(\sum_h x_{ih} \phi_h \right) - \sum_i \exp \left(\sum_h x_{ih} \phi_h \right),$$

where x_{ih} is the ih -th entrie of the model matrix X .

Then, to apply the NR method, we let $g(\phi) = L(\phi)$ and since

$$q_j = \frac{\partial L(\phi)}{\partial \phi_j} = \sum_i n_i x_{ij} - \sum_i m_i x_{ij}$$

$$h_{jk} = \frac{\partial^2 L(\phi)}{\partial \phi_j \partial \phi_k} = - \sum_i m_i x_{ij} x_{ik}$$

so that

$$q_j^{(t)} = \sum_i (n_i - m_i^{(t)}) x_{ij}$$

and

$$h_{jk}^{(t)} = - \sum_i m_i^{(t)} x_{ij} x_{ik}.$$

Now $m^{(t)}$, the t -th approximation for $\hat{m}$, is obtained from $\phi^{(t)}$ through $m^{(t)} = \exp(X\phi^{(t)})$. It generates the next value $\phi^{(t+1)}$ using formula (2.9), which is, for a Poisson model,

$$\phi^{(t+1)} = \phi^{(t)} - [X' \text{Diag}(m^{(t)}) X]^{-1} X'(n - m^{(t)}), \quad (2.10)$$

where $\text{Diag}(m^{(t)})$ has elements $\{m_i^{(t)}\}$ on the main diagonal. Formula (2.10) will produce the third guess and so forth. $\phi^{(t+1)}$ can be expressed alternatively as

$$\phi^{(t+1)} = - [H^{(t)}]^{-1} r^{(t)}, \quad (2.11)$$

where $r^{(t)}$ is the vector with entries

$$r_j^{(t)} = \sum_i m_i^{(t)} x_{ij} \left\{ \log m_i^{(t)} + \frac{(n - m_i^{(t)})}{m_i^{(t)}} \right\}.$$

Note that the expression in braces corresponds to the first term of the Taylor series expansion of $\log n_i$ at $m_i^{(t)}$.

For the initial guess, we take $m_i^{(t)} = n_i$, or $m_i^{(t)} = n_i + 0.5$ if $n_i = 0$. Then equation (2.11) produces the sequence of iterates which can be expressed in a matrix form as

$$\phi^{(t+1)} = (X' \hat{V}^{-1} X)^{-1} X' \hat{V}^{-1} y,$$

where $\hat{V}^{-1} = [\text{Diag}(\hat{m}^{(t)})]^{-1}$ and $y = \log m_i^{(t)} + (n - m_i^{(t)})/m_i^{(t)}$. As t increases, $m^{(t)}$ and $\phi^{(t)}$ usually converge respectively to the MLE $\hat{m}^{(t)}$ and $\hat{\phi}^{(t)}$, respectively. Note that $-H^{-1}$ corresponds to the covariance matrix of $\hat{\phi}$.

Compared to other methods, NR method is usually faster. In fact, for large value of t , the convergence satisfies, for each j ,

$$|\phi_j^{(t+1)} - \hat{\phi}_j| \leq c |\phi_j^{(t)} - \hat{\phi}_j|^2, \text{ for some } c > 0$$

and is referred to as second-order convergence (Agresti, 1990). On the other hand, we have to deal with the inversion of a large matrix when ϕ has high dimension. Moreover, the second derivatives matrix H tend to be complicated for incomplete cases. Thus this method could necessitate careful algebraic manipulations and efficient programming (Little and Rubin, 1987). In those cases, EM algorithm may be a good alternative strategy.

Chapter 3

Modeling the missing data mechanism

3.1 MAR Process

A MAR mechanism has been modeled by Kenward and Molenberghs (1998). Let $\mathbf{Y}_i$ be a vector random variable which contains the complete set of measurements on the i -th individual and let $\mathbf{R}_i$ be an associated vector indicating if each component of $\mathbf{Y}_i$ is missing or not ($i = 1, \dots, n$). For the j -th element of $\mathbf{Y}_i$, we can form a pair $(\mathbf{y}_i, \mathbf{r}_i)$ where $\mathbf{r}_i$ takes the values 1 or 0, indicating, respectively, whether the corresponding value of $\mathbf{y}_i$ is observed or not. In addition, we partition $\mathbf{Y}_i$ into two sets, $(\mathbf{y}_{i,obs}, \mathbf{y}_{i,mis})$ corresponding to observed and missing data respectively. We assume that the joint distribution of $(\mathbf{Y}_i, \mathbf{R}_i)$ respects the customary regularity conditions (e.g. see Cox and Hinkley (1974)).

As defined in section 2.2, the probability of an observation being missing under a MCAR mechanism is independent of the responses, i.e.

$$P(\mathbf{R}_i = \mathbf{r}_i | \mathbf{y}_i) = P(\mathbf{R}_i = \mathbf{r}_i).$$

Therefore, the joint distribution of the observed data under MCAR can be decomposed as follows:

$$f(\mathbf{y}_{i,obs}, \mathbf{r}_i) = f(\mathbf{y}_{i,obs})f(\mathbf{r}_i).$$

Under a MAR mechanism, the probability of an observation being missing is conditionally independent of the unobserved data, given the values of the observed data. That is,

$$P(\mathbf{R}_i = \mathbf{r}_i | \mathbf{y}_i) = P(\mathbf{R}_i = \mathbf{r}_i | \mathbf{y}_{i,obs}) = f(\mathbf{r}_i | \mathbf{y}_{i,obs}, \boldsymbol{\beta}).$$

The joint distribution of the observed data for the i -th individual under MAR mechanism can be decomposed as

$$f(\mathbf{y}_{i,obs}, \mathbf{r}_i) = f(\mathbf{y}_{i,obs}; \boldsymbol{\theta}) f(\mathbf{r}_i | \mathbf{y}_{i,obs}; \boldsymbol{\beta}),$$

for parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, based on a sample of n individuals. The objective is to make inference on $\boldsymbol{\theta}$ treating $\boldsymbol{\beta}$ as the nuisance parameter vector. The complete log-likelihood function can be written as

$$\begin{aligned} & \sum_{i=1}^n \log(y_{i,obs}, r_i) \\ &= \sum_{i=1}^n \log f(y_{i,obs}, \boldsymbol{\theta}) + \sum_{i=1}^n \log f(r_i | y_{i,obs}, \boldsymbol{\beta}), \end{aligned}$$

and

$$l(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{y}_{obs}^{(n)}, \mathbf{r}^{(n)}) = l_1(\boldsymbol{\theta}; \mathbf{y}_{obs}^{(n)}) + l_2(\boldsymbol{\beta}; \mathbf{r}^{(n)}). \quad (3.1)$$

where $\mathbf{y}_{obs}^{(n)} = (\mathbf{y}_{1,obs}^T, \dots, \mathbf{y}_{n,obs}^T)^T$, and $\mathbf{r}^{(n)} = (\mathbf{r}_1^T, \dots, \mathbf{r}_n^T)^T$ and l is the complete log-likelihood function, l_1 and l_2 are the partitioned parts of l respectively ignoring the missing data mechanism or not.

Unless otherwise stated, it is assumed that $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are not interdependent in the sense of Rubin (1976): $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are said to be not interdependent, or distinct, if their joint parameter space factorizes into a $\boldsymbol{\theta}$ -space and a $\boldsymbol{\beta}$ -space, and when prior distributions are specified for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, the priors are independent.

We now state conditions under which one can make valid inferences under a MAR mechanism simply using $l_1(\cdot)$ of equation (3.1). First, we note that under MAR

mechanism $\mathbf{r}_i$ is not an ancillary statistic for $\boldsymbol{\theta}$. Consequently, it is not justified to restrict inferences to the first component of the log-likelihood function. Nevertheless, considering the properties of the procedures below, let us define, for this moment, the appropriate sampling distribution suggested by the pair $(\mathbf{y}_i, \mathbf{r}_i)$, i.e. make use of the partitioned log-likelihood function (3.1). We call this the unconditional sampling framework. Hence, working under this framework means considering the missing value mechanism, or considering $l_2(\cdot)$. This is compared with a second sampling framework, called the naive sampling framework. We formulate the latter by fixing $\mathbf{r}$ by design, that is, if we repeatedly sample using the distribution $f(\mathbf{y}_{obs}; \boldsymbol{\theta})$. If this sampling distribution were appropriate, as in the MCAR case, this would lead directly to the use of $l_1(\cdot)$ as a basis for inference.

Let us clarify some points which come out of the log-likelihood function (3.1). First, notice that the MLE obtained by maximizing only the first part of equation (3.1), $l_1(\boldsymbol{\theta}; \mathbf{y}_{obs})$, is identical to that obtained from maximizing the complete log-likelihood function (3.1). In the same way, the MLE of $\boldsymbol{\beta}$ does not depend on $\boldsymbol{\theta}$ and so any ML ratio involving $\boldsymbol{\theta}$, despite of a common $\boldsymbol{\beta}$, will involve $l_1(\cdot)$ only. Therefore, since we get the same statistics whether derived from $l_1(\cdot)$ or the complete log-likelihood, it follows at once that they have the required properties under the naive sampling framework.

An important element of likelihood-based inference is that we can derive measures of precision of the MLE's from the information gotten from the data. For this, either the observed information i_O can be used, where

$$i_O(\theta_j, \theta_k) = \frac{-\partial^2 l(\cdot)}{\partial \theta_j \partial \theta_k},$$

or the expected information i_E , where

$$i_E(\theta_j, \theta_k) = E \{i_O(\theta_j, \theta_k)\}. \quad (3.2)$$

The argument above justifying the use of the MLE's from $l_1(\boldsymbol{\theta}; \mathbf{y}_{obs})$ applies equally well to the use of the inverse of the observed information derived from $l_1(\cdot)$ as an estimate of the asymptotic variance-covariance matrix of these estimators. On the contrary, using the expected information matrix stated in equation (3.2) presents

some problems. In fact, this expectation needs to be taken over the unconditional sampling distribution (the unconditional information i_U). Consequently the use of the naive sampling framework (producing the naive information i_N) can lead to inconsistent estimates of precision, since we solely use the observed information, ignoring the missing data mechanism. Nevertheless, subsection 3.1.1 shows how to calculate the unconditional information by taking the expectations over the appropriate distribution and thus correct for this bias. This correction allows a direct examination of the effect of ignoring the missing value mechanism on the expected information.

3.1.1 Illustration

Suppose that each member of the pair of observations (Y_{i1}, Y_{i2}) , from unit i , $i = 1, \dots, n$, is a binary random variable, with associated probabilities $P(Y_{i1} = 1) = \lambda$ and $P(Y_{i2} = 1) = \theta$. It is assumed that Y_1 is not missing and that a MAR mechanism is operating with respect to the second observation; that is, the probability of Y_{i2} being missing depends on Y_{i1} alone. We want to compare the naive information i_N with the unconditional information i_U for this setup. We begin by assuming that Y_{i1} and Y_{i2} are independent. The joint distribution of Y_{i1} , Y_{i2} and R_i can then be partitioned as follows : $f(y_{i1}, y_{i2}, r_i) = f(y_{i1})f(y_{i2})f(r_i|y_{i1})$. As the joint distribution of observed Y_{i2} is product binomial, we can express the associated likelihood function as:

$$L(\theta) \propto \prod_{i=1}^m (\theta^{y_{i2}} (1 - \theta)^{1-y_{i2}}) = \theta^{\sum_{i=1}^m y_{i2}} (1 - \theta)^{m - \sum_{i=1}^m y_{i2}},$$

where m denotes the number of Y_2 -values observed at the second occasion. Now taking the logarithm of $L(\theta)$, we obtain the log-likelihood $l(\theta)$

$$l(\theta) = \sum_{i=1}^m y_{i2} \ln \theta + (m - \sum_{i=1}^m y_{i2}) \ln(1 - \theta).$$

The observed information on θ is given by

$$-\frac{d^2 l(\theta)}{d\theta^2} = \frac{\sum_{i=1}^m y_{i2}}{\theta^2} + \frac{(m - \sum_{i=1}^m y_{i2})}{(1 - \theta)^2} = i_O(\theta, \theta). \quad (3.3)$$

Hence, from (3.3), one can get the naive information as

$$\begin{aligned}
 i_N(\theta, \theta) &= E\{i_O(\theta, \theta)\} \\
 &= E\left\{\frac{\sum_{i=1}^m y_{i2}}{\theta^2} + \frac{(m - \sum_{i=1}^m y_{i2})}{(1 - \theta)^2}\right\} \\
 &= \frac{1}{\theta^2}(m\theta) + \frac{1}{(1 - \theta)^2}(m - m\theta) \\
 &= \frac{m}{\theta(1 - \theta)}.
 \end{aligned}$$

The unconditional information is derived in two steps. First, we obtain the conditional expectation of (3.3) with respect to $Y|R$. For this we need

$$\begin{aligned}
 E_{Y|R_i=1}(Y_{i2}) &= P(Y_{i2} = 1|r_i = 1) \\
 &= P(Y_{i2} = 1) = \theta,
 \end{aligned}$$

because of the independence of (Y_{i1}, R_i) and Y_{i2} under the MAR mechanism. It follows that

$$E_{Y|R}\{i_O(\theta, \theta)\} = \frac{m}{\theta(1 - \theta)}.$$

We are treating m as the realization of a random variable, M , over which we take expectation to obtain the unconditional expectation. Under independence, the data are MCAR and setting $\pi = P(R_i = 1)$, we have

$$\begin{aligned}
 i_U(\theta, \theta) &= E_R\left(\frac{m}{\theta(1 - \theta)}\right) \\
 &= \frac{E_R(m)}{\theta(1 - \theta)} \\
 &= \frac{n\pi}{\theta(1 - \theta)}.
 \end{aligned}$$

Replacing π by the estimate m/n , it can be seen that in practice the naive and the unconditional information are equivalent. Therefore, sampling based inferences that use the naive information are valid.

We now introduce dependence between Y_{i1} and Y_{i2} . This can be expressed through the conditional success probabilities of Y_{i2} : $\theta_1 = P(Y_{i2} = 1|y_{i1} = 1)$ and $\theta_0 = P(Y_{i2} = 1|y_{i1} = 0)$. Since the off-diagonal elements of the observed information matrix are zero, we need to consider only the information for one of θ_0 and θ_1 to contrast the naive and the unconditional forms of the expected information. We can write the joint distribution of y_{i1}, y_{i2} as

$$f(y_{i1}, y_{i2}) = \lambda^{y_{i1}} \theta_1^{y_{i1}y_{i2}} (1 - \theta_1)^{y_{i1}(1-y_{i2})} (1 - \lambda)^{1-y_{i1}} \theta_0^{(1-y_{i1})y_{i2}} (1 - \theta_0)^{(1-y_{i1})(1-y_{i2})}.$$

Then, the log-likelihood function is given by

$$\begin{aligned} l(\theta_0, \theta_1) &= \sum_{i=1}^m \log f(y_{i1}, y_{i2}) \\ &= \left(\sum y_{i1}y_{i2} \right) \ln \theta_1 + \left(\sum y_{i1}(1-y_{i2}) \right) \ln(1-\theta_1) \\ &\quad + \text{terms involving } \theta_0 \end{aligned} \tag{3.4}$$

Now, the observed information for θ_1 can be easily found from (3.4)

$$-\frac{d^2 l(\theta_0, \theta_1)}{d\theta_1^2} = \frac{1}{\theta_1^2} \sum_{i=1}^m y_{i1}y_{i2} + \frac{1}{(1-\theta_1)^2} \sum_{i=1}^m y_{i1}(1-y_{i2}) = i_O(\theta_1, \theta_1). \tag{3.5}$$

Taking the expectation of (3.5), we get the naive information as

$$i_N(\theta_1, \theta_1) = \frac{m\lambda}{\theta_1(1-\theta_1)}.$$

Now, we would like to get the unconditional information. For this, we need to consider the conditional expectation over $Y|R$ and define $\eta_0 = P(R_i = 1|y_{i1} = 0)$ and $\eta_1 = P(R_i = 1|y_{i1} = 1)$. Hence, we have $P(R_i = 1) = \eta = \lambda\eta_1 + (1-\lambda)\eta_0$. Then, we have

$$\begin{aligned} E_{Y|R}(Y_{i2}) &= P(Y_{i2} = 1|r_i = 1) \\ &= \frac{\lambda\theta_1\eta_1 + (1-\lambda)\theta_0\eta_0}{\lambda\eta_1 + (1-\lambda)\eta_0}, \end{aligned}$$

$$\begin{aligned}
E_{Y|R}(Y_{i1}Y_{i2}) &= P(Y_{i1} = 1, Y_{i2} = 1 | r_i = 1) \\
&= \frac{\lambda\theta_1\eta_1}{\lambda\eta_1 + (1 - \lambda)\eta_0} \\
&= \frac{\lambda\theta_1\eta_1}{\eta},
\end{aligned}$$

and similarly for $E_{Y|R}(Y_{i1}(1 - Y_{i2}))$,

$$\begin{aligned}
E_{Y|R}(Y_{i1}(1 - Y_{i2})) &= E_{Y|R}(Y_{i1}) - E_{Y|R}(Y_{i1}Y_{i2}) \\
&= P(Y_{i1} = 1 | r_i = 1) - \frac{\lambda\theta_1\eta_1}{\eta} \\
&= \frac{\lambda\eta_1}{\eta} - \frac{\lambda\theta_1\eta_1}{\eta} \\
&= \frac{\lambda\theta_1(1 - \theta_1)}{\eta}.
\end{aligned}$$

Using these conditional expectations in (3.5), we get

$$E_{Y|R}\{i_O(\theta_1, \theta_1)\} = \frac{m\lambda\eta_1}{\eta\theta_1(1 - \theta_1)}.$$

Now taking the expectation over R, we have $E_R(m) = n\eta$, and hence

$$E_R \left[E_{Y|R}\{i_O(\theta_1, \theta_1)\} \right] = \frac{n\lambda\eta_1}{\theta_1(1 - \theta_1)} = i_U(\theta_1, \theta_1). \quad (3.6)$$

We find a similar expression for $i_U(\theta_0, \theta_0)$. We can now examine the conditions under which the naive and the unconditional expectations are the same. From (3.5) and (3.6), $E_R(i_N(\theta_1, \theta_1)) = i_U(\theta_1, \theta_1)$ and $E_R(i_N(\theta_0, \theta_0)) = i_U(\theta_0, \theta_0)$ if $E_R(m/n) = \eta_1 = \eta_0$ and thus $\eta = \eta_1 = \eta_0$. In conclusion, if this condition is satisfied, we have an MCAR mechanism of nonresponse. Hence, we can say that MCAR mechanism is both a necessary and sufficient condition for the equivalence of the naive and the unconditional forms of information.

3.2 Family I (MCAR process)

3.2.1 The model

Under MCAR and using notation of Chen and Fienberg (1974), let x_{ij} denote the fully cross-classified count for the (i, j) cell of an $r \times c$ contingency table, R_i ($i = 1, 2, \dots, r$) the count of the partially classified individuals corresponding to the i^{th} row and C_j ($j = 1, 2, \dots, c$) the count of the partially classified individuals corresponding to the j^{th} column. The total sample size is

$$\begin{aligned} N &= \sum_{ij} x_{ij} + \sum_i R_i + \sum_j C_j \\ &= x_{++} + R_+ + C_+. \end{aligned}$$

We assume that $x_{ij} > 0$ for $i = 1, 2, \dots, r$ and $j = 1, 2, \dots, c$.

The Chen-Fienberg method consists in allocating observations to their observed cells in two steps. In the first step, the procedure assigns observations over the two-way cross-classification, given that they have a multinomial or Poisson distribution. We first discuss here the case of a multinomial distribution. Then at the second step, the probabilities of losing its row identity or column identity, say $\lambda_{1(i)}$ or $\lambda_{2(j)}$ respectively; are given to each observation. These latter observations will now be considered as the corresponding row or column supplemental margins, respectively. The model assumes that the probability of an observation being completely missing, i.e. at both the first and the second step, is zero. Also, the probability of an observation belonging to both row and column supplemental margins is zero. A second assumption on the missingness probability is that $\lambda_{1(i)} > 0$ for $i = 1, 2, \dots, r$ and $\lambda_{2(j)} > 0$ for $j = 1, 2, \dots, c$. As a result, an observation stays in its original cell, (i, j) , with probability $1 - \lambda_{1(i)} - \lambda_{2(j)}$. Clearly, in order for this to be respected, we must assume that $1 \geq \lambda_{1(i)} + \lambda_{2(j)}$ for all i and j . Given that the probability of allocation of the (i, j) cell at the first step is π_{ij} , where $\sum_i \sum_j \pi_{ij} = 1$, one can summarize the probability related to cell (i, j) at the second step as in Table 3.1. One can verify that the sum over all eight cells probabilities is 1.

	1	2	Row supplement- tal margin
1	$(1 - \lambda_{1(1)} - \lambda_{2(1)})\pi_{11}$	$(1 - \lambda_{1(1)} - \lambda_{2(2)})\pi_{12}$	$\lambda_{1(1)}\pi_{1+}$
2	$(1 - \lambda_{1(2)} - \lambda_{2(1)})\pi_{21}$	$(1 - \lambda_{1(2)} - \lambda_{2(2)})\pi_{22}$	$\lambda_{1(2)}\pi_{2+}$
Column supple- mental margin	$\lambda_{2(1)}\pi_{+1}$	$\lambda_{2(2)}\pi_{+2}$	

Table 3.1: Underlying probabilities for a 2×2 table with two sets of partially classified margins

If we now consider that the original sampling scheme is a Poisson with expected values $\{m_{ij}\}$ (with $m_{ij} > 0$, for all i, j), then the table summarizing the probability associated with cell (i, j) at the end of the second step looks like Table 3.1, except that the π_{ij} 's are replaced by m_{ij} 's.

Under Poisson sampling, the likelihood function is proportional to

$$\exp(-\sum \sum m_{ij}) \prod_{i=1}^r \prod_{j=1}^c [(1 - \lambda_{1(i)} - \lambda_{2(j)}) m_{ij}]^{x_{ij}} \prod_{i=1}^r [\lambda_{1(i)} m_{i+}]^{R_i} \prod_{j=1}^c [\lambda_{2(j)} m_{+j}]^{C_j}, \quad (3.7)$$

where $m_{i+} = \sum_{j=1}^c m_{ij}$ and $m_{+j} = \sum_{i=1}^r m_{ij}$ for all i and j .

Note that (3.7) is in fact a product of two functions

$$f_1 = \prod_{i=1}^r \prod_{j=1}^c (1 - \lambda_{1(i)} - \lambda_{2(j)})^{x_{ij}} \prod_{i=1}^r \lambda_{1(i)}^{R_i} \prod_{j=1}^c \lambda_{2(j)}^{C_j}, \quad (3.8)$$

$$f_2 = \exp(-\sum \sum m_{ij}) \prod_{i=1}^r \prod_{j=1}^c m_{ij}^{x_{ij}} \prod_{i=1}^r m_{i+}^{R_i} \prod_{j=1}^c m_{+j}^{C_j}. \quad (3.9)$$

It is interesting to notice that (3.8) is a function of λ 's and (3.9) is a function of m_{ij} 's only. Therefore, we can maximize these functions separately. The principal goal of the Chen-Fienberg method is to make inferences about the parameters of the first step, i.e. to make inferences about the cell parameters m_{ij} 's.

Chen and Fienberg considered two models of general interest for two-way tables: (1)

the unrestricted model and (2) the model of independence of variables corresponding to rows and columns, i.e.

$$m_{ij} = \frac{m_{i+}m_{+j}}{m_{++}}, \quad i=1,2,\dots,r, \quad j=1,2,\dots,c.$$

For the purpose of this work, we will concentrate on the former model. It is possible to represent the unrestricted model in a log-linear form as

$$\log m_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \quad i=1,2,\dots,r, \quad j=1,2,\dots,c.$$

where $\sum_{i=1}^r \alpha_i = 0$, $\sum_{j=1}^c \beta_j = 0$ and $\sum_{i=1}^r \gamma_{ij} = \sum_{j=1}^c \gamma_{ij} = 0$. Note that there is a one-to-one correspondence between the set of parameters $\{m_{ij}\}$ and the set of parameters $\{\mu, \alpha_i, \beta_j, \gamma_{ij}\}$ of the log-linear model.

Now, we want to find the ML equations by taking the derivatives of the logarithm of the function f_2 , given by (3.9) with respect to the parameters μ, α_i, β_j and γ_{ij} . We get the following:

$$\begin{aligned} \frac{\partial \log f_2}{\partial \mu} &= -\sum_{i=1}^r \sum_{j=1}^c m_{ij} + \sum_{i=1}^r \sum_{j=1}^c x_{ij} + \sum_{i=1}^r R_i + \sum_{j=1}^c C_j \\ &= -m_{++} + x_{++} + R_+ + C_+, \end{aligned} \quad (3.10)$$

$$\frac{\partial \log f_2}{\partial \alpha_i} = -m_{i+} + x_{i+} + R_i + \sum_k C_k \left(\frac{m_{ik}}{m_{+k}} \right), \quad (3.11)$$

$$\frac{\partial \log f_2}{\partial \beta_j} = -m_{+j} + x_{+j} + \sum_k R_k \left(\frac{m_{kj}}{m_{k+}} \right) + C_j, \quad (3.12)$$

$$\frac{\partial \log f_2}{\partial \gamma_{ij}} = -m_{ij} + x_{ij} + R_i \left(\frac{m_{ij}}{m_{i+}} \right) + C_j \left(\frac{m_{ij}}{m_{+j}} \right). \quad (3.13)$$

Now, set equations (3.10)-(3.13) to zero and solve for m_{ij} to get

$$\hat{m}_{ij} = x_{ij} + R_i \left(\frac{\hat{m}_{ij}}{\hat{m}_{i+}} \right) + C_j \left(\frac{\hat{m}_{ij}}{\hat{m}_{+j}} \right), \quad i=1,2,\dots,r, \quad j=1,2,\dots,c. \quad (3.14)$$

The left hand side of (3.14) is the estimate of cell count in the (i, j) cell after the unobservable first step and the right hand side is the observed count in the (i, j) cell plus a proportional allocation of the supplemental margins.

3.2.2 Iterative process

In general, (3.14) cannot be solved for $\hat{m}_{ij}$ in closed form, but there exists an iterative procedure, inspired by the EM algorithm, which converges (as proved in section 2.3) to the solution. We simply need an initial estimate of the $\hat{m}_{ij}$'s which could be taken as

$$\hat{m}_{ij}^{(0)} = \left(\frac{x_{ij}}{x_{++}} \right) N, \quad (3.15)$$

and iterate over the MLE equation (3.14) as follow

$$\hat{m}_{ij}^{(v+1)} = x_{ij} + R_i \left(\frac{\hat{m}_{ij}^{(v)}}{\hat{m}_{i+}^{(v)}} \right) + C_j \left(\frac{\hat{m}_{ij}^{(v)}}{\hat{m}_{+j}^{(v)}} \right), \quad (3.16)$$

where $\hat{m}_{ij}^{(v+1)}$ represents the $(v+1)^{th}$ step in the iterative process. We continue iterating until the $\hat{m}_{ij}^{(v)}$'s converge to a desired degree of accuracy.

Suppose now the data follow a multinomial distribution. Then, the likelihood function at the second step of the process is proportional to

$$\prod_{i=1}^r \prod_{j=1}^c [(1 - \lambda_{1(i)} - \lambda_{2(j)}) \pi_{ij}]^{x_{ij}} \prod_{i=1}^r [\lambda_{1(i)} \pi_{ij}]^{R_i} \prod_{j=1}^c [\lambda_{2(j)} \pi_{ij}]^{C_j}. \quad (3.17)$$

As in the Poisson sampling scheme model, (3.17) can be written as a product of two functions, one involving λ 's and the second involving only the π 's

$$\begin{aligned} f_1 &= \prod_{i=1}^r \prod_{j=1}^c (1 - \lambda_{1(i)} - \lambda_{2(j)})^{x_{ij}} \prod_{i=1}^r \lambda_{1(i)}^{R_i} \prod_{j=1}^c \lambda_{2(j)}^{C_j}, \\ f_3 &= \prod_{i=1}^r \prod_{j=1}^c \pi_{ij}^{x_{ij}} \prod_{i=1}^r \pi_{i+}^{R_i} \prod_{j=1}^c \pi_{+j}^{C_j}. \end{aligned} \quad (3.18)$$

Since f_3 in equation (3.18) for the multinomial distribution is proportional to f_2 in equation (3.9) for the Poisson distribution, the derived MLE's are also appropriate.

3.2.3 Multinomial sampling scheme related to Poisson distribution

Suppose that a set of observed counts x_{ij} has a Poisson distribution, then the probability mass function (pmf) can be expressed as

$$f(x_{ij}, m_{ij}) = \frac{\exp(-m_{ij})m_{ij}^{x_{ij}}}{x_{ij}!}, \text{ for } x_{ij} = 0, 1, 2, \dots \quad (3.19)$$

where obviously $m_{ij} = E(x_{ij}) = Var(x_{ij})$. This model assumes that $\{x_{ij}\}$ are independent Poisson random variables. As previously stated, the total sample size of fully observed data $n = \sum \sum x_{ij}$, but we must notice that this quantity is random. Suppose now we fix the total sample size n , then we no longer have a Poisson sampling scheme since each x_{ij} cannot be larger than n . Conditional on n , the x_{ij} 's quantity are no longer independent, since the value of one can change the possible value of the other. Given that $\sum \sum x_{ij} = n$, the conditional probability of x_{ij} 's satisfying this condition is

$$\begin{aligned} & P(x_{ij} \text{ observations in cell}(i, j), \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c \mid \sum \sum x_{ij} = n) \\ &= \frac{P(x_{ij} \text{ observations in cell}(i, j), \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c)}{P(\sum \sum x_{ij} = n)} \\ &= \frac{\prod_i \prod_j \frac{\exp(-m_{ij})m_{ij}^{x_{ij}}}{x_{ij}!}}{\frac{\exp(-\sum m_{kl})\sum m_{kl}}{n!}} = \left(\frac{n!}{\prod_i \prod_j x_{ij}!} \right) \prod_i \prod_j \pi_{ij}^{x_{ij}}, \end{aligned}$$

where $\pi_{ij} = m_{ij}/(\sum \sum m_{ij})$. But this corresponds to the multinomial distribution, with sample size n and probability π_{ij} .

Consequently, MLE's from Poisson distribution and multinomial distribution are the same after conditioning on $\{n\}$. This could be proved from the results of Birch (1963) who showed that the marginal totals are equal to the ML estimates of their expectation, i.e. $x_{ij} = \hat{m}_{ij}$. This could be derived using the following. For the sake of simplicity, let $n = (n_1, n_2, n_3, n_4)'$ and $m = (m_1, m_2, m_3, m_4)'$ where $n_1 = n_{11}$, $n_2 = n_{12}$, $n_3 = n_{21}$, $n_4 = n_{22}$, and $m_1 = m_{11}$, $m_2 = m_{12}$, $m_3 = m_{21}$, $m_4 = m_{22}$.

Expression (3.19) could also be expressed as

$$\begin{aligned}
f(x_{ij}, m_i) &= \exp[x_{ij} \log(m_i) - m_i - \log(x_{ij}!)] \\
&= \exp[x_{ij} \theta_i - \exp(\theta_i) - \log(x_{ij}!)],
\end{aligned}$$

where $\theta_i = \log(m_i)$. Using this information, one can build the loglinear form of this model as

$$\log \mathbf{m} = \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{X}$ is the design matrix and $\boldsymbol{\beta}$ is the model parameter vector. We can also write this model in matrix form :

$$\log(\mathbf{m}) = \begin{bmatrix} \log(m_{11}) \\ \log(m_{12}) \\ \log(m_{21}) \\ \log(m_{22}) \end{bmatrix} = \begin{bmatrix} \log(m_1) \\ \log(m_2) \\ \log(m_3) \\ \log(m_4) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \mathbf{X}\boldsymbol{\beta}.$$

A similar equation can be found after conditioning on the total sample size $\sum \sum x_{ij} = n$ of the cell counts; i.e. the conditional distribution of the independent Poisson random variables n_i ($i = 1, 2, 3, 4$) given n is multinomial with parameters $\pi_i = m_i / (\sum \sum m_i)$. Expressing the Poisson loglinear model for expected cell counts m_i 's ($i = 1, 2, 3, 4$), we have

$$\log(m_i) = \alpha + \mathbf{x}_i \boldsymbol{\beta}_r,$$

where $(1, \mathbf{x}_i)$ is the i -th row of the design matrix $\mathbf{X}$ and $(\alpha, \boldsymbol{\beta}_r)'$ is the design parameter vector. Therefore, the Poisson log-likelihood is

$$\begin{aligned}
l(\alpha, \boldsymbol{\beta}_r) &= \sum n_i \log(m_i) - \sum m_i \\
&= \sum n_i (\alpha + \mathbf{x}_i \boldsymbol{\beta}_r) - \sum \exp(\alpha + \mathbf{x}_i \boldsymbol{\beta}_r) \\
&= n\alpha + \sum n_i \mathbf{x}_i \boldsymbol{\beta}_r - \tau,
\end{aligned}$$

where $\tau = \sum m_i = \sum \exp(\alpha + \mathbf{x}_i \boldsymbol{\beta}_r)$. Since $\log(\tau) = \alpha + \log[\sum \exp(\mathbf{x}_i \boldsymbol{\beta}_r)]$, we can rewrite the log-likelihood as

$$l(\tau, \beta_r) = \left\{ \sum n_i(\mathbf{x}_i \beta_r) - n \log \left[\sum \exp(\mathbf{x}_i \beta_r) \right] \right\} + [n \log(\tau) - \tau]. \quad (3.20)$$

Note now that $\pi_i = m_i / (\sum m_i) = \exp(\alpha + \mathbf{x}_i \beta_r) / [\sum \exp(\alpha + \mathbf{x}_i \beta_r)]$, and that $\exp(\alpha)$ cancels in the numerator and denominator. Unconditionally, $n = \sum n_i$ has a Poisson distribution with mean $\sum m_i = \tau$, so $[n \log(\tau) - \tau]$, which is the second part of the log-likelihood expressed in equation (3.20), is the Poisson log-likelihood for n . Consequently, note that the Poisson loglinear model requires one parameter more than the multinomial loglinear model.

Now returning to the Chen-Fienberg model, we verify that the previous results can be applied to a partially classified table. More specifically, for the Poisson sampling scheme we have

$$\hat{m}_{++} = x_{++} + R_+ + C_+ = N.$$

Using Birch's results, we can write that the MLE's of π_{ij} under the multinomial distribution are

$$\hat{\pi}_{ij} = \frac{\hat{m}_{ij}}{N}, \quad (3.21)$$

where $\hat{m}_{ij}$ is the corresponding MLE's under the Poisson distribution. Then, using the correspondence given in (3.21), one can use the procedure derived above to find a solution for both $\hat{\pi}_{ij}$'s and $\hat{m}_{ij}$'s.

3.2.4 Variance formulae

Asymptotic variance of $\hat{\pi}_{ij}$ ($i = 1, 2, j = 1, 2$) for a 2×2 table can be found using the information matrix and the fact that the variance of $\hat{\pi}_{ij}$ is equal to the inverse of the information matrix. Then, under the unrestricted model, if we substitute $\hat{\lambda}_1 = R_+/N$ and $\hat{\lambda}_2 = C_+/N$ (which are nothing but the MLE's of λ_1, λ_2 respectively), we have

$$\text{var}(\hat{\pi}_{11}) = \frac{\pi_{11}(1 - \pi_{11})}{N(1 - \lambda_1 - \lambda_2)} - \frac{\lambda_1 \pi_{11}^2 \pi_{2+}}{N_\beta(1 - \lambda_1 - \lambda_2) \pi_{1+}} - \frac{\lambda_2 \pi_{11}^2 \pi_{+2}}{N_\beta(1 - \lambda_1 - \lambda_2) \pi_{+1}}$$

$$- \frac{\lambda_1 \lambda_2}{N_\beta (1 - \lambda_1 - \lambda_2)^2} \left(\frac{\pi_{11} \pi_{12} \pi_{21} \pi_{22}}{\pi_{1+} \pi_{2+} \pi_{+1} \pi_{+2}} \right) \left[\pi_{11} (1 - \pi_{11}) \left(\frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}} \right) - 1 \right], \quad (3.22)$$

where

$$N_\beta = N \left[1 + \frac{\lambda_1 \lambda_2}{(1 - \lambda_1 - \lambda_2)} \left(\frac{\pi_{11} \pi_{12} \pi_{21} \pi_{22}}{\pi_{1+} \pi_{2+} \pi_{+1} \pi_{+2}} \right) \left(\frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}} \right) \right].$$

The asymptotic variance formulae for $\hat{\pi}_{12}$, $\hat{\pi}_{21}$ and $\hat{\pi}_{22}$ can be obtained with formulae similar to (3.22).

3.2.5 Goodness-of-fit tests

Formulae for goodness-of-fit tests are similar to the complete case. Hence, we have

$$X^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(x_{ij} - \hat{\theta}_{ij})^2}{\hat{\theta}_{ij}} + \sum_{i=1}^r \frac{(R_i - \hat{\rho}_i)^2}{\hat{\rho}_i} + \sum_{j=1}^c \frac{(C_j - \hat{\sigma}_j)^2}{\hat{\sigma}_j}, \quad (3.23)$$

and

$$G^2 = 2 \left[\sum_{i=1}^r \sum_{j=1}^c x_{ij} \log \left(\frac{x_{ij}}{\hat{\theta}_{ij}} \right) + \sum_{i=1}^r R_i \log \left(\frac{R_i}{\hat{\rho}_i} \right) + \sum_{j=1}^c C_j \log \left(\frac{C_j}{\hat{\sigma}_j} \right) \right], \quad (3.24)$$

where

$$\begin{aligned} \hat{\theta}_{ij} &= (1 - \hat{\lambda}_1 - \hat{\lambda}_2) \hat{m}_{ij} \\ \hat{\rho}_i &= \hat{\lambda}_1 \hat{m}_{i+} \\ \hat{\sigma}_j &= \hat{\lambda}_2 \hat{m}_{+j}. \end{aligned}$$

Both have an asymptotic χ^2 distribution. The contingency table contains $(r+1)(c+1)-1$ cells with observed data; then for a 2×2 table plus two supplemental margins, we have to estimate (4-1) parameters for $\hat{\theta}$, (2-1) parameter for $\hat{\rho}$ and $\hat{\sigma}$, respectively, and (2-1) parameter for $\hat{\lambda}$. Consequently, we have a χ^2 distribution with 2 degrees of freedom.

3.3 Family II

Stasny (1987) considered some Markov chain models for the process generating non-response. Her model is in fact an improvement of the Chen and Fienberg model because it includes a temporal component, i.e. we observe the result at each period of a longitudinal study allowing individuals to be missing at one or both of the measurements.

Stasny's model consists in classifying respondents into one of K non-overlapping categories over two consecutive periods. Then, it is possible to find three different groups of individuals: (1) a group of individuals who responded at both consecutive periods, (2) a group from whom we have measurements in one but not both periods and (3) a group of individuals for whom we do not have measurements in either period.

According to this pattern, we can summarize this information in a $K \times K$ ($K = 1, 2, \dots, k$) matrix for two consecutive periods, say $t - 1$ and t . Hence, the information on respondents who belongs to the second group of individuals, i.e. those who answered at period $t - 1$ but not at period t , and those who answered at period t but not at period $t - 1$, are classified respectively in a column and row supplement. Individuals who did not respond to either period $t - 1$ and t are grouped in a single "missing" cell, i.e. the cell "M" of the current matrix. This setting is displayed in Table 3.2, where x_{ij} = number of sampled individuals with classification i at time $t - 1$ and j at time t ; R_i = number of individuals who were nonrespondents at time t and had classification i at time $t - 1$; C_j = number of individuals who were nonrespondents at time $t - 1$ and had classification j at time t , and M = number of individuals who were nonrespondents at both time $t - 1$ and t .

As an extension of ML estimation for family I, we deal with two different stages. In the unobserved first stage, individuals are allocated to the cells of a $K \times K$ matrix according to probabilities from a Markov chain. Let π_i be the initial probability that an individual is in state i at time $t - 1$, where $\sum_i \pi_i = 1$ and p_{ij} be the transition probability from state i to state j , where $\sum_j p_{ij} = 1$ for all i .

The second stage of Stasny's model corresponds to the first stage of Chen and Fienberg (1974). Hence, each individual in cell (i, j) may either a) be a respondent at time $t - 1$

Time $t - 1$	Time t				
	1	2	...	K	Row supplement
1	x_{11}	x_{12}	...	x_{1K}	R_1
2	x_{21}	x_{22}	...	x_{2K}	R_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
K	x_{K1}	x_{K2}	...	x_{KK}	R_K
Column Supplement	C_1	C_2	...	C_K	M

Table 3.2: Observed data under Stasny's model

and, then, lose the row classification, b) be a respondent at time t and lose the column classification, or c) be a nonrespondent at both times and lose both the row and the column classification. So let $\xi(i, j)$ be the initial probability that an individual in the (i, j) cell of the matrix responds at time $t - 1$; $\varrho_{RR}(i, j)$ be the transition probability from respondent at time $t - 1$ to respondent at time t , and $\varrho_{RR}(i, j)$ be the transition probability from nonrespondent at time $t - 1$ to nonrespondent at time t . We observe data after this second stage.

3.3.1 The model

The likelihood function for the observed data under the two-stage model is proportional to

$$\begin{aligned}
L(\xi(i, j), \varrho_{RR}(i, j), \pi_i, p_{ij}, \varrho_{MM}) &\propto \left\{ \prod_i \prod_j [\xi(i, j) \varrho_{RR}(i, j) \pi_i p_{ij}]^{x_{ij}} \right\} \\
&\times \left\{ \prod_i \left[\sum_j \xi(i, j) (1 - \varrho_{RR}(i, j)) \pi_i p_{ij} \right]^{R_i} \right\} \\
&\times \left\{ \prod_j \left[\sum_i (1 - \xi(i, j)) (1 - \varrho_{MM}(i, j)) \pi_i p_{ij} \right]^{C_j} \right\} \\
&\times \left\{ \left[\sum_i \sum_j (1 - \xi(i, j)) \varrho_{MM}(i, j) \pi_i p_{ij} \right]^M \right\}, \tag{3.25}
\end{aligned}$$

where i and j take on the values from 1 to K . In order to make inference for this model, we use ML estimation. However, there are in total $4K^2 + K$ parameters with $K+1$ constraints on the parameters in the likelihood function (3.25) and only $(K+1)^2$ cells of observed counts with the single constraint that the observed counts sum to the total sample size. Thus, we have to decrease the number of parameters to be estimated. Using only the parameters which are related to the nonresponse, say ξ and ϱ , we get the following four models:

Model A : $\xi(i, j) = \xi$, $\varrho_{RR}(i, j) = \varrho_{RR}$, $\varrho_{MM}(i, j) = \varrho_{MM}$.

Model B : $\xi(i, j) = \xi(i)$, $\varrho_{RR}(i, j) = \varrho_{RR}$, $\varrho_{MM}(i, j) = \varrho_{MM}$.

Model C : $\xi(i, j) = \xi$, $\varrho_{RR}(i, j) = \varrho_{RR}(i)$, $\varrho_{MM}(i, j) = \varrho_{MM}(i)$.

Model D : $\xi(i, j) = \xi$, $\varrho_{RR}(i, j) = \varrho_{RR}(j)$, $\varrho_{MM}(i, j) = \varrho_{MM}(j)$.

Then, we obtain $2K-2$ and $K-1$ degrees of freedom for models A and B respectively, and models C and D are saturated.

Concerning the probabilities π and p , we can say that under model A, the initial probability of having a measurement for an individual at time $t - 1$ is the same for all survey classifications. In the same way, p from respondent to respondent or from nonrespondent to nonrespondent does not depend on the survey classification. Under model B, π depends on the individual's classification while p does not. Under models C and D, the behavior of π is the same as under model A. However, p depends on the survey classification at time $t - 1$ under model C while it depends on the classification at time t under model D.

As done for equation (3.7), the likelihood function (3.25) under model A can be split into two different functions, one involving only the π and p parameters, and a second one involving the nonresponse parameters, ξ and ϱ . Consequently, estimates of π and p can be found separately from the estimates of ξ and ϱ . In terms of Markov chains, this means that we will have two separate and independent Markov chains to estimate the parameters, say one for the measurement process and another one for nonresponse process.

Similarly, under model B, the likelihood function can be written as the product of two functions: one involving π , p and ξ another one involving only the ϱ parameters. According to these two functions and since ξ depends on the measurement process,

one can derive a single Markov chain for the measurement process and three for nonresponse.

In model C, we have dependence of the measurement process through the ϱ parameters. Then, we will separate the likelihood function into two, with one involving parameters π , p and ϱ_{MM} and a second one involving the remaining parameters, ξ and ϱ_{RR} . Consequently, we obtain a single Markov chain for measurement process and three others for nonresponse. Thus, parameters π , p and ϱ_{MM} can be estimated separately from those obtained for ξ and ϱ_{RR} .

Finally, under model D the dependence is the same as in model C and then the two resulting likelihood functions are the same except that the second one involve only ξ .

We also observe the same number of Markov chain for the two different processes.

We maximize the likelihood functions involving π and p under models A, B, C and D using Lagrange multipliers and imposing that $\sum_i \pi_i = 1$ and $\sum_j p_{ij} = 1$ for all i . In general, one must use iterative procedure in order to find the desired estimates. Closed-form solution and iterative equations are given in Appendix B.

3.3.2 Goodness-of-fit tests

After fitting the data to each model, the cells probabilities may be estimated using the formulae in Table 3.3. We simply multiply these probabilities to N , the total sample size, in order to get the expected cell counts required to compute goodness-of-fit tests. Formulae similar to (3.23) and (3.24) may be used, replacing appropriate values of expected cell counts.

Either the Pearson statistic X^2 or the likelihood ratio statistic, G^2 , can be compared to a χ^2 distribution with the appropriate degrees of freedom stated above to see how models A and B fit the data. Models C and D are saturated and hence fit the data perfectly.

Time $t - 1$	Time t				Row supplement
	1	2	...	K	
1					
2					
$\vdots$	$\xi(i, j)\varrho_{RR}(i, j)\pi_i p_{ij}$				$\sum_j \xi(i, j)(1 - \varrho_{RR}(i, j))\pi_i p_{ij}$
K					
Column Supplement	$\sum_i (1 - \xi(i, j))(1 - \varrho_{MM}(i, j))\pi_i p_{ij}$				$\sum_i \sum_j (1 - \xi(i, j))\varrho_{MM}(i, j)\pi_i p_{ij}$

Table 3.3: Expected cell counts for observed data under Stasny's model

3.4 Family III

Family III arises from Fay (1986) and is explained in subsection 3.4.2. In order to have a better understanding of this family, the reader needs to know more about Goodman's theory about loglinear models, which is presented in the following subsection.

3.4.1 Goodman's model

Family III is inspired by Goodman's (1970, 1972, 1973) causal modeling approach. It presents loglinear models in their multiplicative form. For example, let us define four variables A, B, C, and D. Suppose that the relationship between these four variables is described by a $2 \times 2 \times 2 \times 2$ table of counts, where the count in cell (h, i, j, k) is denoted by f_{hijk} , $h = 1, 2$, $i = 1, 2$, $j = 1, 2$, and $k = 1, 2$.

Let us now consider the causal system described by the diagram of Figure 3.1. The five lines in this diagram pertain to the relationship between the following pairs of variables: (1) variables A and B, (2) variables A and C, (3) variables B and C, (4) variables B and D, and (5) variables C and D. Note that there is no line in Figure 3.1 corresponding to the relationship between variables A and D. These five relationships can be described by five two-way tables of counts for each pair of variables. These are called marginal tables of counts and may be obtained from the complete table of

Figure 3.1: Diagram of the causal system among variables A, B, C, and D

counts by summing over the other two variables not included in the pair. For example, $\sum_{j=1}^2 \sum_{k=1}^2 f_{hijk} = f_{hi++}^{AB}$ corresponds to the observed frequency (h, i) of the marginal table $\{AB\}$. Following Goodman (1972), we also use F_{hijk} to denote the expected frequency in cell (h, i, j, k) of the four-way table of counts. Then, by definition of the observed complete table of counts,

$$\sum_{h=1}^2 \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^2 f_{hijk} = n,$$

and

$$\sum_{h=1}^2 \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^2 F_{hijk} = n, \quad (3.26)$$

where n is the total sample size.

Now let us return to the causal system described in Figure 3.1. The hypothesis expressed by the diagram may be replaced by stating the set of two-way marginals that is sufficient for estimating the expected distribution of individuals within the four-way table using an IPF algorithm as described in Chapter 2. More explicitly, the F_{hijk} under the hypothesis can be expressed by a loglinear unsaturated model as

$$F_{hijk} = \eta \tau_h^A \tau_i^B \tau_j^C \tau_k^D \tau_{hi}^{AB} \tau_{hj}^{AC} \tau_{ij}^{BC} \tau_{ik}^{BD} \tau_{jk}^{CD}, \quad (3.27)$$

where τ 's are parameters to be estimated and η is a constant that is introduced to insure that the F_{hijk} satisfies condition (3.26). The parameter τ_h^A , for instance, can be viewed as pertaining to the main effect of the variable A on the expected frequency F_{hijk} in cell (h, i, j, k) ; the parameter τ_{hi}^{AB} can be viewed as pertaining to the interaction effect between variables A and B on the F_{hijk} and so on. We can also define each parameter as follows:

$$\tau_1^A = \tau^A, \tau_2^A = 1/\tau^A, \tau_{11}^{AB} = \tau_{22}^{AB} = \tau^{AB}, \tau_{12}^{AB} = \tau_{21}^{AB} = 1/\tau^{AB}, \text{ etc.}$$

We call the main effect parameters as the expected odds and the interaction effect parameters as the expected odds-ratio.

Model (3.27) can also be expressed as a system of simultaneous equations that described how each dependent variable in the system is affected by the other variables. For example, let us suppose that variable D could take the value 1 or 2. Then, the odds that variable D will take on value 1 can be written as

$$\omega_{hij}^{\bar{D}} = f_{hij1}/f_{hij2}.$$

Similarly, we define the following odds:

$$\begin{aligned} \omega_{hi.k}^{\bar{C}} &= f_{hi1k}/f_{hi2k}, \\ \omega_{h.jk}^{\bar{B}} &= f_{h1jk}/f_{h2jk}, \\ \omega_{hij}^{\bar{A}} &= f_{1ijk}/f_{2ijk}. \end{aligned}$$

Corresponding to these observed odds, we now define the following expected odds as

$$\Omega_{hij}^{\bar{D}} = F_{hij1}/F_{hij2}, \quad (3.28)$$

$$\Omega_{hi.k}^{\bar{C}} = F_{hi1k}/F_{hi2k}, \quad (3.29)$$

$$\Omega_{h.jk}^{\bar{B}} = F_{h1jk}/F_{h2jk}, \quad (3.30)$$

$$\Omega_{hij}^{\bar{A}} = F_{1ijk}/F_{2ijk}. \quad (3.31)$$

We now define a new set of parameters in terms of the τ parameters defined in (3.27) as follows:

$$\gamma^{\bar{D}} = (\tau_1^D)^2, \gamma_i^{\bar{B}\bar{D}} = (\tau_{i1}^{BD})^2, \gamma_j^{C\bar{D}} = (\tau_{j1}^{CD})^2. \quad (3.32)$$

Applying now (3.29) to (3.31) and using parameterization (3.32), we obtain from model (3.27) that

$$\Omega_{hij}^{\bar{D}} = \gamma^{\bar{D}} \gamma_i^{B\bar{D}} \gamma_j^{C\bar{D}}. \quad (3.33)$$

Formula (3.33) states that the expected odds $\Omega_{hij}^{\bar{D}}$ pertaining to variable D are dependent upon variables B and C, but not upon variable A. Similarly, from model (3.27), we obtain the following formulae:

$$\Omega_{hi.k}^{\bar{C}} = \gamma^{\bar{C}} \gamma_h^{A\bar{C}} \gamma_i^{B\bar{C}} \gamma_k^{D\bar{C}}, \quad (3.34)$$

$$\Omega_{h.jk}^{\bar{B}} = \gamma^{\bar{B}} \gamma_h^{A\bar{B}} \gamma_j^{C\bar{B}} \gamma_k^{D\bar{B}}, \quad (3.35)$$

$$\Omega_{.ijk}^{\bar{A}} = \gamma^{\bar{A}} \gamma_i^{B\bar{A}} \gamma_j^{C\bar{A}}. \quad (3.36)$$

Parameters $\gamma^{\bar{C}}$, $\gamma_h^{A\bar{C}}$, $\gamma_i^{B\bar{C}}$, etc. are defined similarly as in (3.32).

How to define a causal model

In order to define a causal model, one must first rely on the unsaturated model given, for example, in (3.27). Then, several "sub-models" may be expressed as a representation of the system of equations (3.33) to (3.36). For example, a diagram of the causal system described by equations (3.33) and (3.35) is presented in Figure 3.2. But, equations (3.33) and (3.35) are equivalent to the unsaturated model stated in (3.27). Consequently (3.27) is equivalent to Figure 3.2. This is also true for any diagram of the causal system described by any system of equations presented in (3.33) to (3.36). However, in order to distinguish among the various causal systems, we must rely upon additional information pertaining to causality within the system of variables. In addition, note that by making certain assumptions such as which variables are prior to which other variables, one can reduce the number of possible causal systems corresponding to the unsaturated model.

Note now that although there is no line between A and B in Figure 3.2, the model describing this causal system (model (3.27)), includes the parameter τ^{AB} . To further explain this point, we will consider what might happen in a simpler context, say when the system contains only three dichotomous variables: A, B, and C. Consider now the following unsaturated model :

Figure 3.2: Causal diagram for equations (3.33) and (3.35)

$$F_{ijk} = \eta \tau_h^A \tau_i^B \tau_j^C \tau_{hi}^{AB} \tau_{hj}^{AC} \tau_{ij}^{BC}. \quad (3.37)$$

Then, applying the notation used earlier, the expected odds pertaining to variable C is expressed as follows:

$$\Omega_{hi,k}^{\bar{C}} = \gamma^{\bar{C}} \gamma_h^{A\bar{C}} \gamma_i^{B\bar{C}}. \quad (3.38)$$

Similar to the four variables model, the model described by (3.37) is equivalent to the model described by (3.38).

Consider now the causal diagram in Figure 3.3 described by the unsaturated model (3.37). Since the model described by formula (3.38) is equivalent to the model described by (3.37), the causal system in Figure 3.3 could also be described by model (3.37). However, it is interesting to note that model (3.37) includes the parameter τ^{AB} in it, but not Figure 3.3. This simply means that, with respect to the causal system in Figure 3.3, if there is statistical independence between variables A and B in the 2×2 marginal table for these two variables, then the numerical value of the τ^{AB} parameter is less than one. One can also say that the τ^{AB} parameter in model (3.37) belongs to the relationship between variables A and B, given the level of variable C. In other words, parameter τ^{AB} belongs to the relationship between variables A and B within the three variables system. It does not belong to the 2×2 marginal table of variables A and B, but the parameter τ^{AB} added to the other parameters τ , affects this relationship.

Figure 3.3: Diagram of the causal system among variables A, B, and C

3.4.2 Fay's model

Fay (1986) considered two nested patterns of non-response under causal models. A pattern of responses is said to be nested or monotone when the variables of interest may be ordered so that complete response is obtained for a set of variables, say u ; response to v_1 , an observed variable, is obtained only for a subset of the respondents; response to v_2 is obtained only for a subset of respondents to v_1 , and so on. Nested patterns can appear by chance or because of the study design, as in, for example, a longitudinal study.

Remember from sections 2.2 and 3.1 that when the response mechanism is ignorable, likelihood inferences based on the observed data are equivalent to the inferences based on the complete likelihood reflecting both the observed data and the response mechanism. The following model for a nested pattern of missing data fulfills the assumption of ignorability.

According to the causal system of each model, Fay developed an EM algorithm to estimate the model parameters. The first step consists in stating the causal system carefully. Then, one needs to determine what are the odds of response for both variables of interest. For a nested pattern of missing data, the odds of response for the second variable, given response on the first variable may be denoted by

$$\Omega_{hi}^{\bar{K}} = F_{hi11}/F_{hi12} = \gamma^{\bar{K}} \gamma_h^{I\bar{K}}. \quad (3.39)$$

Similarly, the odds of response on the first variable are

$$\Omega_{hi}^{\bar{J}} = F_{hi1.}/F_{hi2.} = \gamma^{\bar{J}}. \quad (3.40)$$

The E-step consists in estimating the following complete cross-classifications $f_{hijk}^{(q)}$:

$$f_{hi11}^{(q)} = f_{hi11}, \quad (3.41)$$

$$f_{hi21}^{(q)} = 0, \quad (3.42)$$

$$f_{hi12}^{(q)} = f_{h.12} \frac{\tilde{F}_{hi12}^{(q)}}{\sum_i \tilde{F}_{hi12}^{(q)}}, \quad (3.43)$$

$$f_{hi22}^{(q)} = f_{..22} \frac{\tilde{F}_{hi22}^{(q)}}{\sum_h \sum_i \tilde{F}_{hi22}^{(q)}}, \quad (3.44)$$

where $\tilde{F}_{hijk}^{(q)}$ denotes the provisional estimates of the causal model obtained from the previous step with the following recursive formula:

$$\tilde{F}_{hijk} = \bar{F}_{hijk} \frac{\hat{F}_{hijk}}{\sum_k \hat{F}_{hijk}}, \quad (3.45)$$

where $\bar{F}_{hijk}$ and $\hat{F}_{hijk}$ represent the MLE's (or the extended MLE's) corresponding to $\Omega_{hi}^{\bar{K}}$ and $\Omega_{hi}^{\bar{J}}$ (extended loglinear models) under the causal system, respectively. We use "1" as initial values for $\tilde{F}_{hijk}^{(1)}$, but other values consistent with the causal model may be also employed.

The M-step consists of computing MLE's $\tilde{F}_{hijk}^{(q+1)}$ from the $f_{hijk}^{(q)}$. This requires iterative computation and can be done using an iterative proportional fitting as described in Chapter 2.

The algorithm may converge to different solutions depending on the starting values if the maximum is not unique. One may experiment with different starting values $\tilde{F}_{hijk}^{(1)}$ to investigate the uniqueness of the solution.

3.4.3 Goodness-of-fit test

Goodness-of-fit test is computed with the following equation:

$$G^2 = 2 \sum_{h=1}^2 \sum_{i=1}^2 \sum_{j=1}^2 \sum_{k=1}^2 f_{hijk} \ln \left(\frac{f_{hijk}}{\tilde{F}_{hijk}} \right), \quad (3.46)$$

and can be compared to a χ^2 distribution with the appropriate degrees of freedom. The degrees of freedom are established given the following way: we first have 7

distinct observed cells (since we have a nested pattern of nonresponse). The model presented in this section corresponds to seven identifiable parameters: one for total, three for the marginal distribution and interaction of H and I (under the assumption of ignorability), two in (3.39), and one in (3.40).

3.5 Family IV (Composite linear model)

The fourth family is based on a composite linear model (CLM), proposed by Baker (1992), which consists of a matrix model for incomplete multinomial data. CLM may be used for both ignorable and non-ignorable missing data mechanisms. It can also accommodate marginal models and ordered categorical data. CLM is composed of two steps: a first step where we formulate the data set as a CLM and a second step where we invoke a maximization procedure developed according to CLM features.

3.5.1 The model

We use here a model for marginal homogeneity. As an example, we will fit a cumulative logit model to the marginal probabilities. The model is defined for a 3×3 table with row and column supplements. Let π_{ij} denote the joint probability of response i to A and response j to B, where $\sum_{i=1}^3 \sum_{j=1}^3 \pi_{ij} = 1$. Let $p_{1+} = \pi_{1+}$ and $p_{2+} = \pi_{1+} + \pi_{2+}$ denote the cumulative marginal probabilities for the A margin, and let $p_{+1} = \pi_{+1}$ and $p_{+2} = \pi_{+1} + \pi_{+2}$ denote the cumulative marginal probabilities for the B margin. The cumulative logit model has the following form :

$$\begin{aligned} \text{logit}(p_{1+}) &= \theta_1^{(1)}, \\ \text{logit}(p_{2+}) &= \theta_2^{(1)}, \\ \text{logit}(p_{+1}) &= \theta_1^{(1)} + \theta_{shift}^{(1)}, \\ \text{logit}(p_{+2}) &= \theta_2^{(1)} + \theta_{shift}^{(1)}, \end{aligned} \tag{3.47}$$

where $\theta_{shift}^{(1)} = 0$ under marginal homogeneity.

In order to specify all eight independent multinomial probabilities, we model four arbitrarily selected interior probabilities in addition to the four cumulative marginal probabilities :

$$\text{logit}(\pi_{ij}) = \theta_{ij}^{(1)}, \quad \text{for } i = 1, 2, \text{ and } j = 1, 2. \quad (3.48)$$

We obtain the remaining interior probabilities using functions of the cumulative marginal probabilities:

$$\begin{aligned} \pi_{1+} &= p_{1+} & \pi_{2+} &= p_{2+} - p_{1+} \\ \pi_{+1} &= p_{+1} & \pi_{+2} &= p_{+2} - p_{+1}, \end{aligned}$$

we get

$$\begin{aligned} \pi_{13} &= \pi_{11} + \pi_{12} + \pi_{12} \\ \pi_{23} &= \pi_{2+} + \pi_{21} + \pi_{22} \\ \pi_{31} &= \pi_{+1} + \pi_{11} + \pi_{21} \\ \pi_{13} &= \pi_{+2} + \pi_{12} + \pi_{22}, \end{aligned}$$

and finally we write the remaining interior cell probability as one minus the sum of the other interior cell probabilities :

$$\pi_{33} = 1 - (\pi_{11} + \pi_{12} + \pi_{13} + \pi_{21} + \pi_{22} + \pi_{23} + \pi_{31} + \pi_{32}).$$

We will now extend this model to the missing data mechanism. Let α_{ij} (β_{ij}) denote the probability that only A (B) is observed, given i and j . Considering an ignorable pattern of missing data (because the probability an observation is missing depends only on the observed data) and using a logit form to ensure that estimates lie in the interval $[0, 1]$:

$$\begin{aligned} \text{logit}(\alpha_{ij}) &= \theta_A^{(2)}, \\ \text{logit}(\beta_{ij}) &= \theta_B^{(2)}. \end{aligned} \quad (3.49)$$

These two different parameters allow the model to differ on the rate of missingness associated with A and B. Then, we formulate the log-likelihood function. Let $y_{ij}^{(AB)}$, $y_i^{(A)}$, $y_j^{(B)}$ denote the counts associated with the observation of both A and B, only A, and only B, respectively. The kernel of the likelihood is

$$L(\theta^{(1)}, \theta^{(2)}) = \prod_{i=1}^2 \prod_{j=1}^2 (\pi_{ij}(1 - \alpha_{ij} - \beta_{ij}))^{y_{ij}^{(AB)}} \prod_{j=1}^2 \left(\sum_{i=1}^2 \pi_{ij} \alpha_{ij} \right)^{y_i^{(A)}} \prod_{i=1}^2 \left(\sum_{j=1}^2 \pi_{ij} \beta_{ij} \right)^{y_j^{(B)}}, \quad (3.50)$$

where $\pi_{ij} = \pi_{ij}(\theta^{(1)})$, $\alpha_{ij} = \alpha_{ij}(\theta^{(2)})$, $\beta_{ij} = \beta_{ij}(\theta^{(2)})$, $\theta^{(1)}$ is a vector of the parameter (3.47) and (3.48), and $\theta^{(2)}$ is a vector of the parameters in (3.49).

The purpose of the CLM procedure is to simplify maximization of the log-likelihood (3.50). Formulation of CLM necessitates specification of many matrices and functions, but this is easier than writing out each term of the log-likelihood. In addition, matrix design is closer to multiple regression theory and this simplifies computations. Finally, maximization procedure associated with CLM has good convergence properties.

Let Y denote a column vector of the observed counts and U denote a column vector of the corresponding expected counts. For this example, let us consider that Y follows a multinomial distribution. The expected counts for incomplete (observed) data are summarized in C , the composite link matrix, where $U = CU^*$ and U^* is a column vector of expected counts for the complete data. Then, we model U^* as N , a column vector of constants, multiplied by a probability or a product of conditional probabilities. The CLM formulation for the expected counts is

$$U^* = N \cdot \exp \left(\sum_k Q^{(k)}(\theta^{(k)}) \right), \quad (3.51)$$

where $\cdot$ denotes element-by-element multiplication, k indexes the sub-model and $\theta^{(k)}$ is a set of parameters associated with the k -th sub-model. In this example, sub-model 1 refers to the response probability and sub-model 2 refers to the probabilities of observing both A and B, only A, or only B, as defined in (3.47) and (3.49), respectively. We formulate the product as a sum in order to simplify computation of the derivatives. Lastly, we call $Q^{(k)}$ a q -component, with the following form :

$$Q^{(k)} = q^{(k)} \left(W^{(k)} + G^{(k)} H^{(k)} \right),$$

where $q^{(k)}$ is the transformation function, $G^{(k)}$ is the primary transformation matrix, $W^{(k)}$ is the auxiliary transformation matrix, and $H^{(k)}$ is the H-component. The H-component has the following form :

$$H^{(k)} = h^{(k)} \left(Z^{(k)}, X^{(k)} \theta^{(k)} \right),$$

where $h^{(k)}$ is the design function, $X^{(k)}$ is the primary design matrix and $Z^{(k)}$ is the auxiliary design matrix. In fact, the Q-component is a composite of the linear model in the H-component, from which the name of the procedure CLM.

The reader find the following association to have a better understanding of the Q- and H- components. Suppose that we have a Markov chain, then the H-component is considered as the transition probability matrix and the Q-component is used to transform the logarithms of the H-component into the logarithms of multinomial probabilities. Briefly, the H-component models the data and the Q-component adapts the H-component for multinomial probabilities (for example).

3.5.2 Illustration

Let us first introduce the notation that is used with CLM. Let $Y_{15 \times 1} = vcat(Y_{9 \times 1}^{(A,B)}, Y_{3 \times 1}^{(A)}, Y_{3 \times 1}^{(B)})$ where $vcat$ denotes the vertical concatenation, $Y_{9 \times 1}^{(A,B)}$ are the counts in which both A and B are observed, $Y_{3 \times 1}^{(A)}$ are the counts in which only A is observed, and $Y_{3 \times 1}^{(B)}$ are the counts in which only B is observed. Let $U_{15 \times 1} = vcat(U_{9 \times 1}^{(A,B)}, U_{3 \times 1}^{(A)}, U_{3 \times 1}^{(B)})$ denote the expected counts for the incomplete data, and let $U_{27 \times 1}^* = vcat(U_{9 \times 1}^{*(A,B)}, U_{9 \times 1}^{*(A)}, U_{9 \times 1}^{*(B)})$ denote the expected counts for the complete data. The expected counts for the incomplete data are related to those for the complete data by $U_{15 \times 1} = C_{15 \times 27} U_{27 \times 1}^*$, which can be expand as

$$\begin{bmatrix} U_{9 \times 1}^{(A,B)} \\ U_{3 \times 1}^{(A)} \\ U_{3 \times 1}^{(B)} \end{bmatrix} = \begin{bmatrix} I_9 & 0_{3 \times 9} & 0_{3 \times 9} \\ 0_{3 \times 9} & C_{3 \times 9}^{(A)} & 0_{3 \times 9} \\ 0_{3 \times 9} & 0_{3 \times 9} & C_{3 \times 9}^{(B)} \end{bmatrix} \begin{bmatrix} U_{9 \times 1}^{*(A,B)} \\ U_{3 \times 1}^{*(A)} \\ U_{3 \times 1}^{*(B)} \end{bmatrix},$$

where

$$C_{3 \times 9}^{(A)} = \begin{bmatrix} 1_{1 \times 3} & 0_{1 \times 3} & 0_{1 \times 3} \\ 0_{1 \times 3} & 1_{1 \times 3} & 0_{1 \times 3} \\ 0_{1 \times 3} & 0_{1 \times 3} & 1_{1 \times 3} \end{bmatrix},$$

and similarly for $C_{3 \times 9}^{(B)} = [I_3, I_3, I_3]$, I_3 is a 3×3 identity matrix; $1_{1 \times 3}$ a 1×3 vector of 1's, and $0_{1 \times 3}$ a 1×3 vector of 0's.

Then, let $\pi_{9 \times 1} = (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{23}, \pi_{31}, \pi_{32}, \pi_{33})'$, $\alpha_{9 \times 1} = (\alpha_{11}, \alpha_{12}, \alpha_{13}, \alpha_{21}, \alpha_{22}, \alpha_{23}, \alpha_{31}, \alpha_{32}, \alpha_{33})'$, and $\beta_{9 \times 1} = (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{21}, \beta_{22}, \beta_{23}, \beta_{31}, \beta_{32}, \beta_{33})'$. This notation is used when we express the expected counts for the complete data :

$$\begin{aligned} U_{27 \times 1}^* &= N_{27 \times 1} \cdot \Pr(\text{response to A and B}) \\ &= N_{27 \times 1} \cdot \Pr(\text{observing responses to A and B, observing only responses to A,} \\ &\quad \text{or observing only responses to B} \mid \text{responses to A and B}) \\ &= N_{27 \times 1} \cdot \left[\begin{bmatrix} \pi_{9 \times 1} \\ \pi_{9 \times 1} \\ \pi_{9 \times 1} \end{bmatrix} \cdot \begin{bmatrix} 1_{9 \times 1} - \alpha_{9 \times 1} - \beta_{9 \times 1} \\ \alpha_{9 \times 1} \\ \beta_{9 \times 1} \end{bmatrix} \right] \\ &= (n \ 1_{27 \times 1}) \cdot \exp \left[\log \begin{bmatrix} \pi_{9 \times 1} \\ \pi_{9 \times 1} \\ \pi_{9 \times 1} \end{bmatrix} + \log \begin{bmatrix} 1_{9 \times 1} - \alpha_{9 \times 1} - \beta_{9 \times 1} \\ \alpha_{9 \times 1} \\ \beta_{9 \times 1} \end{bmatrix} \right] \\ &= N_{27 \times 1} \cdot \exp [Q_{27 \times 1}^{(1)} + Q_{27 \times 1}^{(2)}], \quad \text{in CLM notation,} \end{aligned}$$

where n is the total sample size. In order to write the H-component for sub-model 1, we use models (3.47) and (3.48):

$$H_{8 \times 1}^{(1)} = (p_{1+}, p_{2+}, p_{+1}, p_{+2}, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})' = \frac{\exp(X_{8 \times 7}^{(1)} \theta_{7 \times 1}^{(1)})}{1 + \exp(X_{8 \times 7}^{(1)} \theta_{7 \times 1}^{(1)})},$$

where

$$\mathbf{X}_{8 \times 7}^{(1)} \boldsymbol{\theta}_{7 \times 1}^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0_{1 \times 4} \\ 0 & 1 & 0 & 0_{1 \times 4} \\ 0 & 0 & 1 & 0_{1 \times 4} \\ 0 & 1 & 1 & 0_{1 \times 4} \\ 0_{4 \times 1} & 0_{4 \times 1} & 0_{4 \times 1} & \mathbf{I}_4 \end{bmatrix} \begin{bmatrix} \theta_1^{(1)} \\ \theta_2^{(1)} \\ \theta_{shift}^{(1)} \\ \theta_{int}^{(1)} \end{bmatrix},$$

where $\boldsymbol{\theta}_{int}^{(1)} = (\theta_{11}^{(1)}, \theta_{12}^{(1)}, \theta_{21}^{(1)}, \theta_{22}^{(1)})'$ is a column vector of nuisance parameters, corresponding to (3.48).

In CLM notation, the Q-component for sub-model 1 corresponds to

$$\mathbf{Q}_{27 \times 1}^{(1)} = \log \left(\mathbf{W}_{27 \times 1}^{(1)} + \mathbf{G}_{27 \times 8}^{(1)} \mathbf{H}_{8 \times 1}^{(1)} \right),$$

where

$$\mathbf{W}_{27 \times 1}^{(1)} = \begin{bmatrix} \mathbf{W}_{9 \times 1}^{\pi-ind} \\ \mathbf{W}_{9 \times 1}^{\pi-ind} \\ \mathbf{W}_{9 \times 1}^{\pi-ind} \end{bmatrix}, \quad \mathbf{G}_{27 \times 8}^{(1)} = \begin{bmatrix} \mathbf{G}_{9 \times 8}^{\pi-cmarg} \\ \mathbf{G}_{9 \times 8}^{\pi-cmarg} \\ \mathbf{G}_{9 \times 8}^{\pi-cmarg} \end{bmatrix}.$$

The matrices $\mathbf{G}_{9 \times 8}^{\pi-cmarg}$ and $\mathbf{W}_{9 \times 1}^{\pi-ind}$ transform $\mathbf{H}_{8 \times 1}^{(1)}$ into $\pi_{9 \times 1}$. To obtain $\mathbf{G}_{9 \times 8}^{\pi-cmarg}$, we write $\mathbf{G}_{9 \times 8}^{\pi-cmarg} = \mathbf{G}_{9 \times 8}^{\pi-ind} \mathbf{G}_{8 \times 8}^{ind-marg} \mathbf{G}_{8 \times 8}^{marg-cmarg}$, where the subscripts cmarg, marg, marg and ind denote cumulative marginal, marginal, and independent cell probabilities, respectively.

$\mathbf{G}_{8 \times 8}^{marg-marg}$ creates a column vector of four marginal and four interior probabilities from $\mathbf{H}^{(1)}$:

$$(\pi_{1+}, \pi_{2+}, \pi_{+1}, \pi_{+2}, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})' = \mathbf{G}_{8 \times 8}^{marg-marg} \mathbf{H}^{(1)},$$

where

$$\mathbf{G}_{8 \times 8}^{marg-cmarg} = \begin{bmatrix} \tilde{G} & 0 & 0 \\ 0 & \tilde{G} & 0 \\ 0 & 0 & \mathbf{I}_4 \end{bmatrix} \quad \text{and} \quad \tilde{G} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

transforms $(p_{1+}, p_{2+})'$ into $(\pi_{1+}, \pi_{2+})'$.

$G_{8 \times 8}^{ind-marg}$ creates a column vector of eight independent cell probabilities from the above column vector of four marginal and four interior probabilities :

$$(\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{23}, \pi_{31}, \pi_{32})' = G_{8 \times 8}^{ind-marg} (\pi_{1+}, \pi_{2+}, \pi_{+1}, \pi_{+2}, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})',$$

where

$$G_{8 \times 8}^{ind-marg} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 1 & 0 & -1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & -1 \end{bmatrix}.$$

$W_{9 \times 1}^{\pi-ind}$ and $G_{9 \times 8}^{\pi-ind}$ creates $\pi_{9 \times 1}$ from the above column vector of eight independent cell probabilities:

$$\pi_{9 \times 1} = W_{9 \times 1}^{\pi-ind} + G_{9 \times 8}^{\pi-ind} (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{23}, \pi_{31}, \pi_{32})',$$

where

$$W_{9 \times 1}^{\pi-ind} = \begin{bmatrix} 0_{8 \times 1} \\ 1 \end{bmatrix} \quad \text{and} \quad G_{9 \times 8}^{\pi-ind} = \begin{bmatrix} I_8 \\ -I_{1 \times 8} \end{bmatrix}.$$

Let us now consider the design matrix for the second sub-model. The H-component is

$$H_{18 \times 1}^{(2)} = \begin{bmatrix} \alpha_{9 \times 1} \\ \beta_{9 \times 1} \end{bmatrix} = \frac{\exp(X_{18 \times \rho}^{(2)} \theta_{\rho \times 1}^{(2)})}{1 + \exp(X_{8 \times \rho}^{(2)} \theta_{\rho \times 1}^{(2)})},$$

where, corresponding to our ignorable model of missing data mechanism,

$$X_{18 \times \rho}^{(2)} \theta_{\rho \times 1}^{(2)} = \begin{bmatrix} 1_{9 \times 1} & 0_{9 \times 1} \\ 0_{9 \times 1} & 1_{9 \times 1} \end{bmatrix} \begin{bmatrix} \theta_A^{(2)} \\ \theta_B^{(2)} \end{bmatrix}.$$

The Q-component for sub-model 2 models the logarithm of the vertical concatenation of the probabilities of observing responses to both A and B ($1_{9 \times 1} - \alpha_{9 \times 1} - \beta_{9 \times 1}$), only responses to A ($\alpha_{9 \times 1}$), and only responses to B ($\beta_{9 \times 1}$) :

$$Q_{27 \times 1}^{(2)} = \log \left(W_{27 \times 1}^{(2)} + G_{27 \times 18}^{(2)} H_{18 \times 1}^{(2)} \right),$$

where

$$W_{27 \times 1}^{(2)} = \begin{bmatrix} 1_{9 \times 1} \\ 0_{9 \times 1} \\ 0_{9 \times 1} \end{bmatrix} \quad \text{and} \quad G_{27 \times 18}^{(2)} = \begin{bmatrix} -I_9 & -I_9 \\ I_9 & 0_{9 \times 9} \\ 0_{9 \times 9} & I_9 \end{bmatrix}.$$

3.5.3 Maximization procedure

The goal of this procedure is to find a point of convergence to equation (3.51). The maximization procedure starts with an EM algorithm. When the EM algorithm begins to converge slowly, the maximization procedure switches to a NR algorithm. The first and the second derivatives are automatically computed using matrices specifying the CLM.

Let us first start with the EM algorithm. Let $\theta = (\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(k)}, \dots)$ denote the vector of all parameters in the model. Using Espeland and Odoroff's (1985) matrix formula, the E-step can be performed by a simple proportional adjustment of the preceding M-step, such as $Y_{15 \times 1}^{(i+1)*} = U_{27 \times 1}^{(i)*}(\theta) \cdot \left((C_{15 \times 27})' \left(Y_{15 \times 1} / \left(C_{15 \times 27} U_{27 \times 1}^{(i)*}(\theta) \right) \right) \right)$, where $'$ denotes element-by-element division and $Y_{15 \times 1}^{(i+1)*}$ is a vector of complete data counts at the $(i+1)^{th}$ iteration. This formula is recognized as the proportional adjustment step of an IPF algorithm. As an initial guess, Baker (1994) recommends to use a 27×1 vector randomly generated from a uniform distribution on the interval $[0, 1]$ as a starting value of $U_{27 \times 1}^*(\theta)$.

As stated above, when each iteration of the EM algorithm increases the likelihood by only a small amount, the maximization procedure switches to the Newton-Raphson algorithm: $\hat{\theta} = \theta + [ObsInf(\theta)]^{-1} Score(\theta)$, where $Score$ is the score vector, and $ObsInf$ is the observed information matrix. As defined in Baker (1994), $Score$ and $ObsInf$ are matrix function of the quantities in the H- and Q-components. Let

$\dot{H}^{(k)} = \dot{h}^{(k)}(Z^{(k)}, X^{(k)}\theta^{(k)})$ and $\ddot{H}^{(k)} = \ddot{h}^{(k)}(Z^{(k)}, X^{(k)}\theta^{(k)})$ denote the first and the second derivatives, respectively, of $H^{(k)}$ with respect to $X^{(k)}\theta^{(k)}$. Also defined $\dot{Q}^{(k)}$ and $\ddot{Q}^{(k)}$ as the first and the second derivatives, respectively, of $Q^{(k)}$ with respect to $G^{(k)}H^{(k)}$. Let $M^{(k)} \equiv G^{(k)} \text{diag}(\dot{H}^{(k)})X^{(k)}$, $S \equiv \text{hcat}(\dot{Q}_i^{(k)}M^{(k)})$. Also let $R \equiv -C'(Y/U)$, for a multinomial likelihood. Then,

$$\text{Score} = -S' \text{diag}(R \cdot U^*)$$

and

$$\begin{aligned} \text{ObsInf} = & \text{blockdiag}|_k X^{(k)'} \text{diag} \left(\ddot{H}^{(k)} G^{(k)'} \left(R \cdot U^* \cdot \dot{Q}^{(k)} \right) \right) X^{(k)} \\ & + \text{blockdiag}|_k M^{(k)'} \text{diag} \left(R \cdot U^* \cdot \dot{Q}^{(k)} \right) M^{(k)} \\ & + S' \text{diag}(R \cdot U^*) S \\ & + S' \text{diag}(U^*) C' \text{diag}(Y/(U \cdot U^*)) C \text{diag}(U^*) S, \end{aligned}$$

where $\text{diag}(\text{vector})$ creates a diagonal matrix with vector on the diagonal and $\text{blockdiag}(\text{matrix1}, \text{matrix2})$ creates a matrix with matrix1 and matrix2 forming a block diagonal.

We will now derive the matrix formula for the score vector and the observed information matrix (Baker (1992, 1994)). Let y_i , c_{ij} , u_i^* denote elements of Y , C and U^* , respectively. When u_i^* contains multinomial probabilities to sum to one, as in our problem, the log-likelihood is:

$$l(\theta) = \sum_{j=1}^n y_i \log \left(\sum_i c_{ij} u_i^*(\theta) \right).$$

Under the CLM, $u_i^* = n_i \exp \left(\sum_k q_i^{(k)} (\omega_i^{(k)}) \right)$, where $\omega_i^{(k)} = \sum_e w_e^{(k)} + g_{ie}^{(k)} h^{(k)}(z_e^{(k)}, x_e^{(k)}\theta)$, n_i is the (i) element of N , $w_e^{(k)}$ is the (e) element of $W^{(k)}$, $g_{ie}^{(k)}$ is the (i, e) element of $G^{(k)}$, $x_e^{(k)}$ is the (e) row of $X^{(k)}$, and $z_e^{(k)}$ is the (e) element of Z .

Let θ_b denote the (b) element in $\theta_{p \times 1} = (\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(K)})$. Let $\delta_b^{(k)} = 1$, if the (b) element of $\theta_{p \times 1}$ is in $\theta^{(k)}$ and 0, otherwise. The (b) element of the first derivative of u_i^* with respect to θ is $\dot{u}_{ib}^* = u_i^* s_{ib}$, where $s_{ib} = \sum_k \dot{q}_i^{(k)} \dot{\omega}_{ib}^{(k)}$, $\dot{\omega}_{ib}^{(k)} = \delta_b^{(k)} \sum_e g_{ie}^{(k)} \dot{h}_e^{(k)} x_{eb}^{(k)}$ and

$\dot{q}_i^{(k)} = \dot{q}(\omega_i^{(k)})$. The (a, b) element of the second derivative of u_i^* with respect to θ is $\ddot{u}_{iab}^* = \ddot{u}_{1ab}^* + \ddot{u}_{2ab}^* + \ddot{u}_{3ab}^*$, where $\ddot{u}_{1ab}^* = u_i^* \sum_k \dot{q}_i^{(k)} \ddot{\omega}_{iab}^{(k)}$, $\ddot{u}_{2ab}^* = u_i^* \sum_k \ddot{q}_i^{(k)} \dot{\omega}_{ia}^{(k)} \dot{\omega}_{ib}^{(k)}$, $\ddot{u}_{3ab}^* = u_i^* s_{ia} s_{ib}$, $\ddot{q}_i^{(k)} = \ddot{q}(\omega_i^{(k)})$, and $\ddot{\omega}_{iab}^{(k)} = \delta_a^{(k)} \delta_b^{(k)} \sum_e g_{ie}^{(k)} \ddot{h}_e^{(k)} x_{ea}^{(k)} x_{eb}^{(k)}$.

Define $u_j \equiv \sum_i c_{ij} u_i^*$ and $r_i \equiv -\sum_j c_{ij} (y_j / u_j)$, then the (b) element of the score vector is

$$\begin{aligned} \text{Score}_b &= -\sum_j \left[\sum_i c_{ij} \dot{u}_{ib}^* \right] (-y_j / u_j) \\ &= -\sum_i r_i \dot{u}_{ib}^*, \quad \text{by definition} \\ &= -\sum_i (r_i u_i^*) s_{ib}, \quad \text{by substitution.} \end{aligned}$$

The (a, b) element of the information matrix is

$$\begin{aligned} \text{ObsInf}_{ab} &= \sum_i r_i \ddot{u}_{iab}^* + \sum_j \left[\sum_i c_{ij} \dot{u}_{ia}^* \right] \left[\sum_i c_{ij} \dot{u}_{ib}^* \right] (y_j / u_j^2) \\ &= \text{ObsInf}_{1ab} + \text{ObsInf}_{2ab} + \text{ObsInf}_{3ab} + \text{ObsInf}_{4ab}, \end{aligned}$$

where

$$\begin{aligned} \text{ObsInf}_{1ab} &= \sum_i r_i \ddot{u}_{1iab}^*, \quad \text{by definition,} \\ &= \sum_i r_i (u_i^* \sum_k \dot{q}_i^{(k)} \ddot{\omega}_{iab}^{(k)}), \quad \text{by substitution,} \\ &= \sum_k \sum_i r_i u_i^* \dot{q}_i^{(k)} \ddot{\omega}_{iab}^{(k)}, \quad \text{by changing the order of summation,} \\ &= \sum_k \sum_i r_i u_i^* \dot{q}_i^{(k)} \delta_a^{(k)} \delta_b^{(k)} \sum_e g_{ie}^{(k)} \ddot{h}_e^{(k)} x_{ea}^{(k)} x_{eb}^{(k)}, \quad \text{by substitution,} \\ &= \sum_k \delta_a^{(k)} \delta_b^{(k)} \left(\sum_e \ddot{h}_e^{(k)} \left(\sum_i g_{ie}^{(k)} (r_i u_i^* \dot{q}_i^{(k)}) \right) \right) x_{ea}^{(k)} x_{eb}^{(k)}, \quad \text{by changing} \\ &\quad \text{the order of summation} \end{aligned}$$

$$\text{ObsInf}_{2ab} = \sum_i r_i \ddot{u}_{2iab}^*, \quad \text{by definition,}$$

$$\begin{aligned}
&= \sum_i r_i (u_i^* \sum_k \ddot{q}_i^{(k)} \dot{\omega}_{ia}^{(k)} \dot{\omega}_{ib}^{(k)}), \quad \text{by substitution,} \\
&= \sum_k \sum_i (u_i^* \ddot{q}_i^{(k)} r_i) \dot{\omega}_{ia}^{(k)} \dot{\omega}_{ib}^{(k)}, \quad \text{by changing the order of summation,}
\end{aligned}$$

$$\begin{aligned}
ObsInf_{3ab} &= \sum_i r_i \ddot{u}_{3iab}^*, \quad \text{by definition,} \\
&= \sum_i r_i u_i^* (s_{ia} s_{ib}), \quad \text{by substitution,}
\end{aligned}$$

$$\begin{aligned}
ObsInf_{4ab} &= - \sum_j \left(\sum_i c_{ij} \dot{u}_{ia}^* \right) \left(\sum_i c_{ij} \dot{u}_{ib}^* \right) (y_j / u_j^2), \\
&= - \sum_j (y_j / u_j^2) \left(\sum_i c'_{ij} (u_i^* s_{ia}) \right) \left(\sum_i c'_{ij} u_i^* s_{ib} \right), \quad \text{by substitution.}
\end{aligned}$$

Finally, let U , S , R , and M denotes matrices or column vectors with element $\{u_i\}$, $\{s_{ib}\}$, $\{r_i\}$, and $\{\omega_i\}$, respectively. The following equations convert the subscript notation into the matrix formulae:

$$\begin{aligned}
Score &= vector1_d vector2_d = (vector1 \cdot vector2)_d \\
ObsInf &= \sum_d matrix1_{de} vector1_d = (matrix1' vector1)_e,
\end{aligned}$$

where $\{vector1_d\}$, $\{vector2_d\}$, $\{matrix1_{de}\}$ and $\{matrix2_{df}\}$ denote elements of vector1, vector2, matrix1 and matrix2, respectively.

3.5.4 CLM and 2×2 tables of counts

In case of 2×2 tables of counts with 2 supplemental margins, it is possible to define a CLM in order to allocate missing data. The model is much simpler. Assuming that the data set contains a "missing at both occasions" cell, represented by $U_{1 \times 1}^{()}$ and $U_{4 \times 1}^{* ()}$, we represent the expected counts for the incomplete data as:

$$\begin{bmatrix} U_{4 \times 1}^{(A,B)} \\ U_{2 \times 1}^{(A)} \\ U_{2 \times 1}^{(B)} \\ U_{1 \times 1}^{(\cdot)} \end{bmatrix} = \begin{bmatrix} I_4 & 0_{2 \times 4} & 0_{2 \times 4} & 0_{1 \times 4} \\ 0_{2 \times 4} & C_{2 \times 4}^{(A)} & 0_{2 \times 4} & 0_{1 \times 4} \\ 0_{2 \times 4} & 0_{2 \times 4} & C_{2 \times 4}^{(B)} & 0_{1 \times 4} \\ 0_{1 \times 4} & 0_{1 \times 4} & 0_{1 \times 4} & 1_{1 \times 4} \end{bmatrix} \begin{bmatrix} U_{4 \times 1}^{*(A,B)} \\ U_{4 \times 1}^{*(A)} \\ U_{4 \times 1}^{*(B)} \\ U_{4 \times 1}^{*(\cdot)} \end{bmatrix}, \quad (3.52)$$

where

$$C_{2 \times 4}^{(A)} = \begin{bmatrix} 1_{1 \times 2} & 0_{1 \times 2} \\ 0_{1 \times 2} & 1_{1 \times 2} \end{bmatrix},$$

and similarly for $C_{2 \times 4}^{(B)} = [I_2, I_2]$, I_2 is a 2×2 identity matrix; $1_{1 \times 2}$ a 1×2 vector of 1's, and $0_{1 \times 2}$ a 1×2 vector of 0's, and similarly for $1_{1 \times 4}$.

The Q- and H-components of a CLM for a multinomial distribution are the following:

$$Q_{16 \times 1} = W_{16 \times 1} + G_{16 \times 16} H_{16 \times 1},$$

where $W_{16 \times 1}$ is a 16×1 matrix of 0's since the model does not require that a transformation of $H_{16 \times 1}$ into the form of the estimates. $G_{16 \times 16}$ is the identity matrix of dimension 16×16 . $H_{16 \times 1}$ represents the 16 estimates of θ , the vector that contains the 16 parameters of the multinomial distribution.

The following method (Baker, 1992) necessitates less quantities to be specified. The expected counts for the incomplete data are defined as in (3.52). Then, one may apply an EM algorithm directly, where the E-step is

$$Y_{12 \times 1}^*(\theta^{(t)}) = U_{12 \times 1}^*(\theta^{(t)}) \cdot \left(C_{8 \times 12}' \left[\frac{Y_{8 \times 1}}{U_{8 \times 1}(\theta^{(t)})} \right] \right),$$

where $\theta^{(t)}$ is the parameter to estimate (may be a vector), and $\cdot$ and the division represent an element-by-element operation as defined in the previous subsection. The M-step consists in computing $U_{12 \times 1}^*(\theta^{(t+1)})$, where $\theta^{(t+1)}$ is the MLE based on $Y_{12 \times 1}^*(\theta^{(t)})$. The M-step is done using an IPF as defined in Chapter 2.

3.5.5 Goodness-of-fit test

The likelihood ratio test G^2 is computed using a formula similar to (3.24). The degrees of freedom are assessed according to the number of cell in matrix U minus one parameter for the total. We subtract to this the number of parameters required in (3.47), (3.48), and (3.49). For example, this correspond to 6 and 2 degrees of freedom for a 3×3 and a 2×2 table of counts, respectively.

3.6 Family V (Closed-form estimates)

Baker, Rosenberg and Dersimonian (1992) considered closed-form estimation for two-way contingency tables with three supplemental margins. They used a loglinear model because of its simplicity and the fact that this kind of model could be easily generalized to multiway tables. Overall, closed-form estimates simplify the fitting of the models.

3.6.1 The model

Define random variables I and J with realizations $i = 1, \dots, n_1$ and $j = 1, \dots, n_2$; where n_1 and n_2 are the number of level for both variables I and J , respectively. As in the other families, the partially observed data set must be square, i.e. $n_1 = n_2$. This requirement is important if one wants to obtain closed-form solutions. Let M_I and M_J be indicators of missing data for I and J , respectively, such that $M_I = 1$ if I is observed and $M_I = 2$ if I is missing. M_J is defined similarly.

Now, using a cross-classification of I , J , M_I , and M_J , the incomplete data for $M_I = M_J = 1$ consists in a table with cell counts denoted by $\{z_{ij11}\}$, and a supplemental I margin when $M_I = 1$ and $M_J = 2$ with cell counts denoted by $\{z_{i+12}\}$. Similarly for a J margin, we have $\{z_{+j21}\}$ and when both variables are missing, i.e. when $M_I = M_J = 2$, we have $\{z_{++22}\}$. Consequently, the complete data set sums to $N = z_{++++}$, the total sample size. This notation is summarized in Table 3.4. In the same manner as for other models, we assume either a Poisson or multinomial distribution. Let $\{\mu_{ijkl}\}$ denote the expected counts for the complete data. One can also rewrite $\{\mu_{ijkl}\}$ in a "u-term" notation, i.e., following a loglinear model for two partially

		$M_J = 1$		$M_J = 2$
		$J = 1$	$J = 2$	
$M_I = 1$	$I = 1$	z_{1111}	z_{1211}	z_{1+12}
	$I = 2$	z_{2111}	z_{2211}	z_{2+12}
$M_I = 2$		z_{+121}	z_{+221}	z_{++22}

Table 3.4: Notation for $n_1 = n_2 = 2$

observed categorical variables, assuming that we do not have any three- or four-way interaction:

$$\begin{aligned} \log(\mu_{ijkl}) = & u + u_I(i) + u_J(j) + u_{M_I}(k) + u_{M_J}(l) + u_{IJ}(ij) + u_{IM_I}(ik) + u_{IM_J}(il) \\ & + u_{JM_I}(jk) + u_{JM_J}(jl) + u_{M_IM_J}(kl), \end{aligned} \quad (3.53)$$

with the constraint that the sum of each u -term over each of the indicated subscripts is zero.

Exponentiating (3.53), we get another parameterization of the model which will lead to the closed-form ML estimates. It is given by

$$\begin{aligned} m_{ij} &= \exp\{u + u_I(i) + u_J(j) + u_{M_I}(1) + u_{M_J}(1) + u_{IJ}(ij) + u_{IM_I}(i1) + u_{IM_J}(i1) \\ &\quad + u_{JM_I}(j1) + u_{JM_J}(j1) + u_{M_IM_J}(11)\}, \\ a_{ij} &= \exp\{-2[u_{M_I}(1) + u_{IM_I}(i1) + u_{JM_I}(j1) + u_{M_IM_J}(11)]\}, \\ b_{ij} &= \exp\{-2[u_{M_J}(1) + u_{IM_J}(i1) + u_{JM_J}(j1) + u_{M_IM_J}(11)]\}, \\ g &= \exp\{4u_{M_IM_J}(11)\}, \end{aligned} \quad (3.54)$$

then we obtain parameters of Table 3.5, where $m_{ij} \geq 0$, $a_{ij} \geq 0$, $b_{ij} \geq 0$, $g \geq 0$, and $\sum_i \sum_j m_{ij} (1 + a_{ij} + b_{ij} + a_{ij}b_{ij}g) = N$. Then, these parameters have the following interpretation:

$$m_{ij} = N \Pr(I = i, J = j, M_I = 1, M_J = 1),$$

$$\begin{aligned}\mu_{ij11} &= m_{ij}, & \mu_{ij12} &= m_{ij}b_{ij}, \\ \mu_{ij21} &= m_{ij}a_{ij}, & \mu_{ij22} &= m_{ij}a_{ij}b_{ij}g,\end{aligned}$$

Table 3.5: Parameterization according to model 3.54

$$\begin{aligned}a_{ij} &= \frac{\Pr(M_I = 2, M_J = 1 \mid I = i, J = j)}{\Pr(M_I = 1, M_J = 1 \mid I = i, J = j)}, \\ b_{ij} &= \frac{\Pr(M_I = 1, M_J = 2 \mid I = i, J = j)}{\Pr(M_I = 1, M_J = 1 \mid I = i, J = j)}, \\ g &= \frac{\Pr(M_I = 1, M_J = 1 \mid I = i, J = j) \Pr(M_I = 2, M_J = 2 \mid I = i, J = j)}{\Pr(M_I = 1, M_J = 2 \mid I = i, J = j) \Pr(M_I = 2, M_J = 1 \mid I = i, J = j)}.\end{aligned}$$

Because we assume that there is no three- and four-way interaction, we conclude that there is no dependence between i and j , so g can be rewritten as

$$g = \frac{\Pr(M_I = 1, M_J = 1) \Pr(M_I = 2, M_J = 2)}{\Pr(M_I = 1, M_J = 2) \Pr(M_I = 2, M_J = 1)}.$$

Notice that if $g = 1$, it implies that the missing-data mechanism is independent.

It is now possible to express some probability equations. For example, the marginal probability for I and J is defined as

$$\pi_{ij..} = \Pr(I = i, J = j) = \frac{m_{ij}(1 + a_{ij} + b_{ij} + a_{ij}b_{ij}g)}{N};$$

as well as the probability of missing in I conditional on not missing in J :

$$\psi_{ij} = \Pr(M_I = 2 \mid M_J = 1, I = i, J = j) = \frac{a_{ij}}{(1 + a_{ij})}.$$

Similarly for the probability of missing in J conditional on not missing in I , we have

$$\gamma_{ij} = \Pr(M_J = 2 \mid M_I = 1, I = i, J = j) = \frac{b_{ij}}{(1 + b_{ij})}.$$

3.6.2 Goodness-of-fit test

Finally, one can derive a formula for the goodness-of-fit test, according to the general equation for G^2 test, and by taking into account the current notation :

$$\begin{aligned}
G^2 = & -2 \sum_i \sum_j z_{ij11} \log \left[\frac{\hat{m}_{ij}}{z_{ij11}} \right] + \sum_i z_{i+12} \log \left[\frac{\sum_j \hat{m}_{ij} \hat{b}_{ij}}{z_{i+12}} \right] \\
& + \sum_j z_{+j21} \log \left[\frac{\sum_i \hat{m}_{ij} \hat{a}_{ij}}{z_{+j21}} \right] + z_{ij22} \log \left[\frac{\sum_i \sum_j \hat{m}_{ij} \hat{a}_{ij} \hat{b}_{ij} \hat{g}}{z_{ij22}} \right],
\end{aligned}$$

where $\hat{m}_{ij}$, $\hat{a}_{ij}$, $\hat{b}_{ij}$ and $\hat{g}$ are the estimates of m_{ij} , a_{ij} , b_{ij} and g , respectively. G^2 has a χ^2 distribution with $(n_1 + n_2 - 2)$ degrees of freedom for model 1 (see subsection 3.6.3) and $(n_1 - 1)$ degrees of freedom for models 2 to 5. Models 6 to 9 fit the data perfectly and then have 0 degree of freedom. The degrees of freedom are derived in the following way: we have 9 observed cells, then $(9-1)$ degrees of freedom. We then subtract the number of parameters to be estimated by each model. For example, in model 1 (see Table 3.7), we have $(4-1)$ parameters for $\hat{m}_{ij}$, plus 1 parameter for $\hat{\alpha}_{..}$, $\hat{\beta}_{..}$, and g : 3 in total. Then, for $n_1 = n_2 = 2$, we subtract 6 parameters; consequently, we have 2 degrees of freedom, which corresponds to $n_1 + n_2 - 2$. Similarly for the other models.

3.6.3 Estimates

We now face the same problem as in family II: the above parameterization involves more parameters than degrees of freedom for estimation, i.e. this loglinear model is not identifiable. However, if $n_1 = n_2$, nine models are identifiable when (given m_{ij} and g) a_{ij} equals either $\alpha_{..}$, $\alpha_{i.}$, or $\alpha_{.j}$ and b_{ij} equals either $\beta_{..}$, $\beta_{i.}$, or $\beta_{.j}$. The subscript '.' indicates the parameter is constant over the corresponding index. α and β are symbols associated with missing data mechanism and represent the sub-categories of a_{ij} and b_{ij} , respectively.

As an example, we derive the first and the fifth model. The others can be obtained similarly given that the MSS's under the complete data according to various models are as shown in Table 3.6.

The first model is the only model where we do not have, in fact, a close-form solution. However, numerical estimates may be found using an EM algorithm. The iterative procedure is represented in Table 3.7. Another important characteristic of this model

#	Model	Meaning	MSS's under complete data
1	$(\alpha_{..}, \beta_{..})$	missing in each variable is constant	z_{++kl}, z_{ij++}
2	$(\alpha_{..}, \beta_{i.})$	missing in one variable is constant and missing in the other variable depends on the realization of the former variable	$z_{++kl}, z_{ij++}, z_{i++l}$
3	$(\alpha_{.j}, \beta_{..})$	missing in one variable is constant and missing in the other variable depends on the realization of the former variable	$z_{++kl}, z_{ij++}, z_{+jk+}$
4	$(\alpha_{..}, \beta_{.j})$	missing in one variable is constant and missing in the other depends on its realization	$z_{++kl}, z_{ij++}, z_{+j+l}$
5	$(\alpha_{i.}, \beta_{..})$	missing in one variable is constant and missing in the other depends on its realization	$z_{++kl}, z_{ij++}, z_{i+k+}$
6	$(\alpha_{i.}, \beta_{i.})$	missing in each variable depends on the realization of the same variable	$z_{++kl}, z_{ij++}, z_{i+k+}, z_{i++l}$
7	$(\alpha_{.j}, \beta_{.j})$	missing in each variable depends on the realization of the same variable	$z_{++kl}, z_{ij++}, z_{+jk+}, z_{+j+l}$
8	$(\alpha_{i.}, \beta_{.j})$	missing in each variable depends on its realization	$z_{++kl}, z_{ij++}, z_{i+k+}, z_{+j+l}$
9	$(\alpha_{.j}, \beta_{i.})$	missing in each variable depends on the realization of the other variable	$z_{++kl}, z_{ij++}, z_{+jk+}, z_{i++l}$

Table 3.6: Meaning of the model and MSS's under the complete data

is that it represents an MCAR missing-data mechanism. This is explained by the fact that the probability that an observation is missing does not depend on the unobserved realization of a random variable. The MSS's under the complete data are z_{++kl} and z_{ij++} . Then, we find the following likelihood equations for z_{++kl} :

$$E(z_{++11}) = \sum_i \sum_j \mu_{ij++} = \hat{m}_{++} = z_{++11}, \quad (3.55)$$

$$E(z_{++12}) = \sum_i \sum_j \mu_{ij12} = \hat{m}_{++} \hat{\beta}_{..} = z_{++12}, \quad (3.56)$$

$$E(z_{++21}) = \sum_i \sum_j \mu_{ij21} = \hat{m}_{++} \hat{\alpha}_{..} = z_{++21}, \quad (3.57)$$

$$E(z_{++22}) = \sum_i \sum_j \mu_{ij22} = \sum_i \hat{m}_{i+} \hat{\alpha}_{..} \hat{\beta}_{..} \hat{g} = z_{++22}. \quad (3.58)$$

From these equations, we get

$$\hat{\alpha}_{..} = \frac{z_{++21}}{z_{++11}}, \quad \hat{\beta}_{..} = \frac{z_{++12}}{z_{++11}}, \quad g = \frac{z_{++11} z_{++22}}{z_{++21} z_{++12}}.$$

Similarly, the likelihood equations for z_{ij++} are

$$\begin{aligned} E(z_{ij++}) &= \sum_k \sum_l \mu_{ijkl} = z_{ij++} \\ &= \mu_{ij11} + \mu_{ij12} + \mu_{ij21} + \mu_{ij22} \\ &= m_{ij} + m_{ij} b_{ij} + m_{ij} a_{ij} + m_{ij} a_{ij} b_{ij} g \\ &= \hat{m}_{ij} (1 + \hat{\alpha}_{..} + \hat{\beta}_{..}) = z_{ij11} + z_{+j21} \frac{\hat{m}_{ij}}{\hat{m}_{+j}} + z_{i+12} \frac{\hat{m}_{ij}}{\hat{m}_{i+}}, \end{aligned} \quad (3.59)$$

after substituting (3.58). Substitution of $\hat{\alpha}_{..}$ and $\hat{\beta}_{..}$ in (3.59) leads to an iterative process which is detailed in Table 3.7.

For the fifth model, the MSS's are z_{++kl} , z_{ij++} and z_{i+k+} . Consequently, the likelihood equations corresponding to the first MSS are:

$$\begin{aligned} E(z_{++11}) &= \sum_i \sum_j \mu_{ij11} = \hat{m}_{++} = z_{++11}, \\ E(z_{++12}) &= \sum_i \sum_j \mu_{ij12} = \hat{m}_{++} \hat{\beta}_{..} = z_{++12}, \\ E(z_{++21}) &= \sum_i \sum_j \mu_{ij21} = \sum_i \hat{m}_{i+} \hat{\alpha}_{i.} = z_{++21}, \\ E(z_{++22}) &= \sum_i \sum_j \mu_{ij22} = \sum_i \hat{m}_{i+} \hat{\alpha}_{..} \hat{\beta}_{..} \hat{g} = z_{++22}. \end{aligned}$$

Then, $\hat{\beta}_{..}$ and g are the same as in model 1. For the MSS z_{i+k+} , the likelihood equations are:

$$E(z_{i+k+}) = \sum_j \sum_l \mu_{ijkl} = z_{i+k+}$$

Model $(\alpha_{..}, \beta_{..})$ d.f. = $n_1 + n_2 - 2$	
$\hat{m}_{ij}^{(0)} = z_{ij11}$ $\hat{m}_{ij}^{(t+1)} = \frac{\left[z_{ij++} + z_{i+12} \frac{\hat{m}_{ij}^{(t)}}{\hat{m}_{i+}} + z_{+j21} \frac{\hat{m}_{ij}^{(t)}}{\hat{m}_{+j}} \right] z_{++11}}{z_{++11} + z_{++12} + z_{++21}}$	
$\hat{\alpha}_{..} = \frac{z_{++21}}{z_{++11}}$ $\hat{\beta}_{..} = \frac{z_{++12}}{z_{++11}}$ $\hat{g} = \frac{z_{++11} z_{++22}}{z_{++12} z_{++21}}$	
Model $(\alpha_{..}, \beta_{.j})$ d.f. = $n_1 - 1$	Model $(\alpha_{.j}, \beta_{..})$ d.f. = $n_2 - 1$
$\hat{m}_{ij} = \frac{z_{ij11} z_{+j+1} z_{++11}}{z_{+j11} z_{++++1}}$ $\hat{\alpha}_{..} = \frac{z_{++21}}{z_{++11}}$ $\hat{\beta}_{..} = \frac{z_{++12}}{\hat{m}_{i+}}$ $\hat{g} = \frac{z_{++11} z_{++22}}{z_{++21} z_{++12}}$	$\hat{m}_{ij} = \frac{z_{ij11} z_{i+1+} z_{++11}}{z_{i+11} z_{++++1}}$ $\hat{\alpha}_{.j} = \frac{z_{+j21}}{\hat{m}_{+j}}$ $\hat{\beta}_{..} = \frac{z_{++12}}{z_{++11}}$ $\hat{g} = \frac{z_{++11} z_{++22}}{z_{++21} z_{++12}}$
Model $(\alpha_{..}, \beta_{.j})$ d.f. = $n_1 - 1$	Model $(\alpha_{i.}, \beta_{..})$ d.f. = $n_2 - 1$
$\hat{m}_{ij} = \frac{z_{ij11} z_{+j+1} z_{++11}}{z_{+j11} z_{++++1}}$ $\hat{\alpha}_{..} = \frac{z_{++21}}{z_{++11}}$ $\hat{\beta}_{.j}$ such that $\sum_j \hat{m}_{ij} \hat{\beta}_{.j} = z_{i+12}$ $\hat{g} = \frac{z_{++11} z_{++22}}{z_{++12} z_{++21}}$	$\hat{m}_{ij} = \frac{z_{ij11} z_{i+1+} z_{++11}}{z_{i+11} z_{++++1}}$ $\hat{\alpha}_{i.}$ such that $\sum_i \hat{m}_{ij} \hat{\alpha}_{i.} = z_{+j21}$ $\hat{\beta}_{..} = \frac{z_{++12}}{z_{++11}}$ $\hat{g} = \frac{z_{++11} z_{++22}}{z_{++12} z_{++21}}$
Model $(\alpha_{..}, \beta_{.j})$ d.f. = 0	Model $(\alpha_{i.}, \beta_{..})$ d.f. = 0
$\hat{m}_{ij} = z_{ij11}$ $\hat{\alpha}_{i.}$ such that $\sum_i z_{ij11} \hat{\alpha}_{i.} = z_{+j21}$ $\hat{\beta}_{.j} = \frac{z_{i+12}}{z_{i+11}}$ $\hat{g} = \frac{z_{++22}}{\sum_i \hat{m}_{i+} \hat{\alpha}_{i.} \hat{\beta}_{.j}}$	$\hat{m}_{ij} = z_{ij11}$ $\hat{\alpha}_{.j} = \frac{z_{+j21}}{z_{+j11}}$ $\hat{\beta}_{.j}$ such that $\sum_j z_{ij11} \hat{\beta}_{.j} = z_{i+12}$ $\hat{g} = \frac{z_{++22}}{\sum_j \hat{m}_{+j} \hat{\alpha}_{.j} \hat{\beta}_{.j}}$
Model $(\alpha_{.j}, \beta_{i.})$ d.f. = 0	
$\hat{m}_{ij} = z_{ij11}$ $\hat{\alpha}_{.j} = \frac{z_{+j21}}{z_{+j11}}$ $\hat{\beta}_{i.} = \frac{z_{i+12}}{z_{i+11}}$ $\hat{g} = \frac{z_{++22}}{\sum_i \sum_j \hat{m}_{ij} \hat{\alpha}_{.j} \hat{\beta}_{i.}}$	
Model $(\alpha_{..}, \beta_{.j})$ d.f. = 0	
$\hat{m}_{ij} = z_{ij11}$ $\hat{\alpha}_{i.}$ such that $\sum_i z_{ij11} \hat{\alpha}_{i.} = z_{+j21}$ $\hat{\beta}_{i.} = \frac{z_{i+12}}{z_{i+11}}$ $\hat{g} = \frac{z_{++22}}{\sum_i \hat{m}_{i+} \hat{\alpha}_{i.} \hat{\beta}_{i.}}$	

Table 3.7: Closed-form models

$$\begin{aligned}
&= \mu_{i111} + \mu_{i112} + \mu_{i211} + \mu_{i212}, \text{ (when } k = 1) \\
&= m_{i1} + m_{i1}b_{i1} + m_{i2} + m_{i2}b_{i2} \\
&= \hat{m}_{i+}(1 + \hat{\beta}_{..}), \text{ according to the model;}
\end{aligned} \tag{3.60}$$

and

$$\begin{aligned}
E(z_{i+k+}) &= \sum_j \sum_l \mu_{ijkl} = z_{i+k+} \\
&= \mu_{i121} + \mu_{i122} + \mu_{i221} + \mu_{i222}, \text{ (when } k = 2) \\
&= m_{i1}a_{i1} + m_{i1}a_{i1}b_{i1}g + m_{i2}a_{i2} + m_{i2}a_{i2}b_{i2}g \\
&= \hat{m}_{i+}\hat{\alpha}_{i.} + \hat{m}_{i+}\hat{\alpha}_{i.}\hat{\beta}_{..}g, \text{ according to the model} \\
&= \hat{m}_{i+}\hat{\alpha}_{i.} + \hat{m}_{i+}\hat{\alpha}_{i.}\frac{z_{++12}}{z_{++11}}\frac{z_{++22}}{z_{++12}z_{++21}}, \text{ replacing } \hat{\beta}_{..} \text{ and } g \\
&= \hat{m}_{i+}\hat{\alpha}_{i.} + \sum_j z_{++22} \frac{\hat{m}_{ij}\hat{\alpha}_{i.}}{\sum_i \hat{m}_{i+}\hat{\alpha}_{i.}}.
\end{aligned}$$

Isolating $\hat{m}_{i+}\hat{\alpha}_{i.}$ and using algebraic manipulations, we obtain:

$$\hat{m}_{i+}\hat{\alpha}_{i.} = \sum_j z_{++21} \frac{\hat{m}_{ij}\hat{\alpha}_{i.}}{\sum_i \hat{m}_{i+}\hat{\alpha}_{i.}}. \tag{3.61}$$

Now, substituting the value of $\hat{\beta}_{..}$ in (3.60) we get

$$\begin{aligned}
\hat{m}_{i+} \left(1 + \frac{z_{++12}}{z_{++11}} \right) &= z_{i+1+} \\
\hat{m}_{i+} \left(\frac{z_{++1+}}{z_{++11}} \right) &= z_{i+1+} \\
\hat{m}_{i+} &= \frac{z_{i+1+}z_{++11}}{z_{++1+}},
\end{aligned}$$

which is useful to find the result of the following likelihood equation.

Finally, we know that the likelihood equation corresponding to z_{ij++} is

$$E(z_{ij++}) = \sum_k \sum_l \mu_{ijkl} = z_{ij++}$$

$$\begin{aligned}
&= \mu_{ij11} + \mu_{ij12} + \mu_{ij21} + \mu_{ij22} \\
&= m_{ij} + m_{ij}b_{ij} + m_{ij}a_{ij} + m_{ij}a_{ij}b_{ij}g \\
&= \hat{m}_{ij} (1 + \hat{\alpha}_{i.} + \hat{\beta}_{..}) = z_{ij11} + z_{+j21} \frac{\hat{m}_{ij}\hat{\alpha}_{i.}}{\sum_i \hat{m}_{ij}\hat{\alpha}_{i.}} + z_{i+12} \frac{\hat{m}_{ij}}{\hat{m}_{i+}}. \quad (3.62)
\end{aligned}$$

From (3.61), we know that the interior solution requires that $\sum_i \hat{m}_{ij}\hat{\alpha}_{i.} = z_{+j21}$. Substituting this, the value of $\hat{\beta}_{..}$ and g into (3.62), we have

$$\begin{aligned}
\hat{m}_{ij} + \hat{m}_{ij}\hat{\alpha}_{i.} + \hat{m}_{ij} \frac{z_{++12}}{z_{++11}} &= z_{ij11} + z_{+j21} \frac{\hat{m}_{ij}\hat{\alpha}_{i.}}{z_{+j21}} + \hat{m}_{ij} \frac{z_{i+12}z_{++1+}}{z_{i+1+}z_{++11}} \\
\hat{m}_{ij} \left(1 + \frac{z_{++12}}{z_{++11}} - \frac{z_{i+12}z_{++1+}}{z_{i+1+}z_{++11}} \right) &= z_{ij11} \\
\frac{z_{ij11}z_{++11}z_{i+1+}}{z_{++1+}(z_{i+1+} - z_{i+12})} &= \hat{m}_{ij},
\end{aligned}$$

which leads to

$$m_{ij} = \frac{z_{ij11}z_{i+1+}z_{++11}}{z_{i+11}z_{++1+}}.$$

Other models are derived similarly based on their different MSS's. Table 3.7 summarizes the closed-form estimates for the various models.

3.6.4 Boundary estimates

One of the requirements of the closed-form estimate is that all solutions to the system of equations must be non-negative. A negative solution means that the ML estimate lies on the boundary of the parameter space. In this case, we need to compute boundary solutions. However, and particularly with large sample sizes, a negative solution means that the fit is poor. In this situation, computation of boundary solutions is useless. In order to get a boundary solution, one simply needs to set a negative parameter estimate to zero in the likelihood equations. An example of boundary solutions may be found in Table 3.8. For example, in model $(\alpha_{i.}, \beta_{..})$, substituting $\hat{\alpha}_{2.} = 0$ into the likelihood equations of section 5.2 leads to the same estimates for $\hat{\beta}_{..}$, $\hat{g}$, and $\hat{m}_{i+}$, since they do not depend on $\hat{\alpha}_{i.}$. Then, replacing

$\hat{\alpha}_{1.} = 0$	$\hat{\alpha}_{2.} = 0$
Model($\alpha_{i.}, \beta_{..}$)	
$\hat{m}_{1j} = \frac{z_{1j11}z_{1+1+}z_{++11}}{z_{1+11}z_{++1+}}$	$\hat{m}_{1j} = \frac{(z_{++11}z_{1+1+})(z_{1j11}+z_{+j21})}{z_{++1+}(z_{1+11}+z_{++21})}$
$\hat{m}_{2j} = \frac{(z_{++11}z_{2+1+})(z_{2j11}+z_{+j21})}{z_{1+1+}(z_{2+11}+z_{++21})}$	$\hat{m}_{2j} = \frac{z_{2j11}z_{2+1+}z_{++11}}{z_{2+11}z_{++1+}}$
$\hat{\alpha}_{2.} = \frac{z_{++21}z_{++1+}}{z_{++11}z_{2+1+}}$	$\hat{\alpha}_{1.} = \frac{z_{++21}z_{++1+}}{z_{++11}z_{2+1+}}$
$\hat{\beta}_{..} = \frac{z_{++12}}{z_{++11}}$	$\hat{\beta}_{..} = \frac{z_{++12}}{z_{++11}}$
$\hat{g} = \frac{z_{++11}z_{++22}}{z_{++12}z_{++21}}$	$\hat{g} = \frac{z_{++11}z_{++22}}{z_{++12}z_{++21}}$
Model($\alpha_{i.}, \beta_{i.}$)	
$\hat{m}_{1j} = z_{1j11}$	$\hat{m}_{1j} = \frac{z_{1+11}(z_{1j11}+z_{+j21})}{(z_{1+11}+z_{++21})}$
$\hat{m}_{2j} = \frac{z_{2+11}(z_{2j11}+z_{+j21})}{(z_{2+11}+z_{++21})}$	$\hat{m}_{2j} = z_{2j11}$
$\hat{\alpha}_{2.} = \frac{z_{++21}}{z_{2+11}}$	$\hat{\alpha}_{1.} = \frac{z_{++21}}{z_{1+11}}$
$\hat{\beta}_{.j} \text{ such that } \sum_j \hat{m}_{ij11}\hat{\beta}_{.j} = z_{i+12}$	$\hat{\beta}_{.j} \text{ such that } \sum_j \hat{m}_{ij11}\hat{\beta}_{.j} = z_{i+12}$
$\hat{g} = \frac{z_{2+11}z_{++22}}{z_{2+12}z_{++21}}$	$\hat{g} = \frac{z_{1+11}z_{++22}}{z_{1+12}z_{++21}}$
Model($\alpha_{i.}, \beta_{.j}$)	
$\hat{m}_{1j} = z_{1j11}$	$\hat{m}_{1j} = \frac{z_{1+11}(z_{1j11}+z_{+j21})}{(z_{1+11}+z_{++21})}$
$\hat{m}_{2j} = \frac{z_{2+11}(z_{2j11}+z_{+j21})}{(z_{2+11}+z_{++21})}$	$\hat{m}_{2j} = z_{2j11}$
$\hat{\alpha}_{2.} = \frac{z_{++21}}{z_{2+11}}$	$\hat{\alpha}_{1.} = \frac{z_{++21}}{z_{1+11}}$
$\hat{\beta}_{.j} \text{ such that } \sum_j \hat{m}_{ij11}\hat{\beta}_{.j} = z_{i+12}$	$\hat{\beta}_{.j} \text{ such that } \sum_j \hat{m}_{ij11}\hat{\beta}_{.j} = z_{i+12}$
$\hat{g} = \frac{z_{2+11}z_{++22}}{z_{2+12}z_{++21}}$	$\hat{g} = \frac{z_{1+11}z_{++22}}{z_{1+12}z_{++21}}$

Table 3.8: Boundary solutions

$\hat{\alpha}_{2.} = 0$ into (3.61), we easily obtain that $\hat{\alpha}_{1.} = (z_{++21}z_{++1+})/(z_{++11}z_{1+1+})$. Using all those estimates and because we already know from section 5.2 that

$$\hat{m}_{ij}(1 + \hat{\alpha}_{i.} + \hat{\beta}_{..}) = z_{ij11} + z_{+j21} \frac{\hat{m}_{ij}\hat{\alpha}_{i.}}{\sum_i \hat{m}_{ij}\hat{\alpha}_{i.}} + z_{i+12} \frac{\hat{m}_{ij}}{\hat{m}_{i+}},$$

when $i = 1$, we find that

$$\begin{aligned}
\hat{m}_{1j} \left(1 + \frac{z_{++21}z_{++1+}}{z_{++11}z_{1+1+}} + \frac{z_{++12}}{z_{++11}} \right) &= z_{1j11} + z_{+j21} \frac{\hat{m}_{1j}\hat{\alpha}_{1.}}{\sum_i \hat{m}_j\hat{\alpha}_i} \\
&\quad + \hat{m}_{1j} \frac{z_{1+12}z_{++1+}}{z_{1+1+}z_{++11}} \\
\hat{m}_{1j} \left(\frac{z_{++1+}z_{1+1+} + z_{++21}z_{++1+} - z_{1+12}z_{++1+}}{z_{++11}z_{1+1+}} \right) &= z_{1j11} + z_{+j21} \\
\hat{m}_{1j} &= \frac{(z_{1j11} + z_{+j21})(z_{++11}z_{1+1+})}{z_{++1+}(z_{1+11} + z_{++21})}.
\end{aligned}$$

Other boundary solutions are found in a similar way.

Chapter 4

Examples

4.1 Family I (Chen and Fienberg model)

For family I, we use a data set analyzed in Little (1993). This example consists of a 2×2 table with supplemental row and column margins. This table arises from an artificial data set. It has two variables, X_1 and X_2 , with J and K levels, $J = K = 2$. It also includes two supplemental margins of dimension $(J \times 1)$ and $(1 \times K)$ to represent respectively the supplemental column and row margins. Data are reproduced in Table 4.1. Parameters estimates (where μ_{ij} represents the cell probability), standard errors, expected data counts, OR, and G^2 and their associated degrees of freedom are presented in Table 4.2 by application of the theory of Chapter 3.

Chen and Fienberg (1974) model has been coded in S-Plus (version 4.5) in Appendix A.1 using formulae (3.15) and (3.16). In order to avoid convergence problems due to lack of precision when we apply a recursive algorithm, we simply designed the algorithm in a way to allow the user to stop the algorithm when the desired level of convergence is obtained. Our results agree with those of Molenberghs (1999a). For this example, the algorithm converges in 16 iterations. Computation of the standard errors and G^2 statistics is possible using formulae (3.22) and (3.24), respectively. Interpretation and analysis of the results may be found in Chen and Fienberg (1974).

	X ₁		missing
X ₂	100	50	30
	75	75	60
missing	28	60	

Table 4.1: Two-way contingency table with two supplemental margin

Parameter	I (MCAR)	Expected count
μ_{11}	0.2790 (0.0234)	83.70
μ_{12}	0.1740 (0.0206)	52.20
μ_{21}	0.2387 (0.0228)	71.61
μ_{22}	0.3078 (0.0242)	92.34
OR	2.0706 (0.3878)	
G ²	9.2512	
d.f.	2	

Table 4.2: Parameter estimates (standard errors) and expected data counts for fitted model to the data in Table 4.1

4.2 Family II (Stasny's model)

We illustrate Stasny's (1987) theory with her own example taken from The Canadian Labour Force Survey (LFS) with approximately 56,000 households. A detailed description of the LFS can be found in the *Methodology of the Canadian Labour Force Survey 1976* (Statistics Canada, 1977). This survey is a monthly rotating panel survey of labor force in Canada and contains 5 periods in total from August 1979 to January 1980. We illustrate here the first period of the survey. The observed LFS data for August to September 1979 are displayed in Table 4.3, based on the overlap sample.

We use the algorithm of Appendix A.2 in order to get the following results. For the same reason as in family I, we designed the algorithm in a way to allow the user to stop the algorithm when the desired level of convergence is obtained. Most of our results agree with Stasny's results, except for slight discrepancies in $\hat{\xi}$ and in the

August 1979	September 1979			
	E	U	N	Row supplement
E	9222	128	662	473
U	221	322	151	59
N	256	164	5941	292
Column supplement	996	69	676	4353

Table 4.3: Observed LFS data for August to September 1979

		Model A		
		E	U	N
Expected cell counts	E	9218.5252	128.8470	660.2522
	U	228.8329	335.8296	155.9852
	N	255.5975	164.8864	5918.2439
Estimates of the π 's	$\hat{\pi}_E$		0.5683	
	$\hat{\pi}_U$		0.0422	
	$\hat{\pi}_N$		0.3714	
Estimates of the p_{ij} 's	E	0.9211	0.0129	0.0660
	U	0.3175	0.4660	0.2165
	N	0.0403	0.0260	0.9337

Table 4.4: Estimates of cell counts, π 's, and transition probabilities p_{ij} 's for August to September 1979

expected data counts. Models A converges quickly with 6 iterations. The procedure of convergence for model B, C and D converge relatively slowly, requiring 210, 91 and 101 iterations, respectively. The estimates of the expected cells counts, the estimates of the π 's and of the transition probabilities, p_{ij} values of Model A are present in Table 4.4. The parameter estimates and G^2 statistic of Model A are given in Table 4.5.

	$\hat{\xi}$	$\hat{\theta}_{RR}$	$\hat{\theta}_{MM}$	G^2	d.f.
Model A	0.7459	0.9539	0.7143	16.45	4

Table 4.5: Parameter estimates and G^2 statistics for August to September 1979

Physical status		1		2	
Mental status		1	2	1	2
Missing (physical)	Missing (mental)				
No	No	13	7	7	16
No	Yes	14			4
Yes	No			0	
Yes	Yes			10	

Table 4.6: Data on mental and physical status

4.3 Family III (Fay's model)

Fay's (1986) model is explained in detail using the following example. Table 4.6 summarizes the data analyzed by Fuchs (1982) on survival of subjects cross-classified by initial evaluations of physical and mental status.

The causal system that we would like to model is the following: nonresponse to physical status is independent of true mental or physical status, and all subjects with missing data on physical status are further not evaluated on mental status. Evaluation of mental status for those with complete physical status are not obtained for some subjects as a consequence of their mental, but not physical, status. This causal system is represented by Figure 4.1, where v_1 and v_2 represent physical (H) and mental (I) status ($h, i = 1$ for poor, $h, i = 2$ for good), respectively; and r_1 and r_2 represent response ($j, k = 1$, response; $j, k = 2$, nonresponse) to physical (J) and mental (K) status, respectively.

The assumption of ignorability of response in this context requires: (1) that obtaining a response for mental status, given that the value of physical status is obtained, depends only on the value of physical status, and (2) that obtaining a response on

Figure 4.1: Example of an estimable model of nonresponse : relationship between physical and mental status

physical status is independent of the unobserved values of physical and mental status. In fact, this is what we have in the causal system described by Figure 4.1.

According to the path diagram of Figure 4.1, the odds for mental status, given response on physical status are denoted by $\Omega_{hi}^{\bar{K}} = \gamma^{\bar{K}} \gamma_i^{I\bar{K}}$. Using Goodman's theory, this is equivalent to the following extended loglinear model:

$$F_{hijk} = \eta \tau_h^H \tau_i^I \tau_j^J \tau_k^K \tau_{hi}^{HI} \tau_{hj}^{HJ} \tau_{ij}^{IJ} \tau_{ik}^{IK} \tau_{jk}^{JK} \tau_{hij}^{HIJ}. \quad (4.1)$$

Similarly, the odds for physical status are $\Omega_{hi}^{\bar{J}} = \gamma^{\bar{J}}$ and are equivalent to the following extended loglinear model:

$$F_{hij.} = \eta \tau_h^H \tau_i^I \tau_j^J \tau_{hi}^{HI}. \quad (4.2)$$

Now, using the EM algorithm described in Chapter 3 for Fay's model and arranging data as in Table 4.6, $f_{hijk}^{(1)}$ from (3.42) to (3.44) are

Physical status		1		2	
Mental status		1	2	1	2
Missing (physical)	Missing (mental)				
No	No	13.00	7.00	7.00	16.00
No	Yes	7.00	7.00	7.00	7.00
Yes	No			0	
Yes	Yes	2.50	2.50	2.50	2.50

If we run an IPF algorithm on the extended loglinear model (4.2), we obtain $\hat{F}_{hijk}^{(2)}$:

Physical status		1		2	
Mental status		1	2	1	2
Missing (physical)	Missing (mental)				
No	No	13.79	10.06	6.21	12.94
No	Yes	6.21	3.94	2.79	5.06
Yes	Yes	2.50	2.50	2.50	2.50

Now, we collapse Table 4.6 over missing on mental status. If we run an IPF algorithm on this new table according to the extended loglinear model (4.1), we obtain $\bar{F}_{hijk}^{(2)}$:

Physical status		1		2	
Mental status		1	2	1	2
Missing (physical)	Missing (mental)				
No	–	19.33	14.18	9.88	17.61
Yes	–	3.17	2.32	1.62	2.89

From (3.45), we obtain $\tilde{F}_{hijk}^{(2)}$:

Physical status		1		2	
Mental status		1	2	1	2
Missing (physical)	Missing (mental)				
No	No	13.33	10.19	6.82	12.66
No	Yes	6.00	3.99	3.06	4.95
Yes	Yes	3.17	2.32	1.62	2.89

The second cycle of the EM algorithm gives us the second e-step with $f_{hijk}^{(2)}$:

Physical status		1		2	
Mental status		1	2	1	2
Missing (physical)	Missing (mental)				
No	No	13.00	7.00	7.00	16.00
No	Yes	8.41	5.59	1.53	2.47
Yes	No		0		
Yes	Yes	3.17	2.32	1.62	2.89

giving $F_{hijk}^{(3)}$ by reapplication of (3.42) to (3.44) :

Physical status		1		2	
Mental status		1	2	1	2
Missing (physical)	Missing (mental)				
No	No	14.11	9.48	5.83	13.59
No	Yes	7.01	3.33	2.89	4.76
Yes	Yes	3.46	2.10	1.43	3.01

The algorithm eventually converges to $F_{hijk}^{(\infty)}$:

Physical status		1		2	
Mental status		1	2	1	2
Missing (physical)	Missing (mental)				
No	No	14.21	7.00	5.79	16.00
No	Yes	12.79	0.00	5.21	0.00
Yes	Yes	4.43	1.15	1.80	2.62

Our results perfectly agree with those from Fay (1986). Note that all answers are rounded after the second decimal. Using the code in Appendix A.3, this algorithm converges after 114 iterations. Using equation (3.46), the likelihood ratio test G^2 is 0.7604 with 0 degree of freedom.

Smoker	Birth weight (in grams)		
	< 2500	$\geq$ 2500	missing
yes	4 512	21 009	1 049
no	3 394	24 132	1 135
missing	142	464	1 224

Table 4.7: Birth weight and smoking: observed counts

4.4 Family IV (Baker's model)

In order to fit a CLM to the data, Baker implemented with Mathematica (version 2.2) several general models that can be modified to fit the desired data set (the code is available upon request to the author). These examples may be used for either a 2×2 or a 3×3 table of counts with supplemental margins. We used one of this example to fit data of the TRICC study in the following chapter.

4.5 Family V (Baker's model)

Family V example considers the data in Table 4.7 from a prospective study of pregnant women to assess the relationship between perinatal factor (smoker or non-smoker) and the subsequent development (weight) in the offspring. Table 4.7 presents mother's self-reported smoking status with newborn's weight (< 2500 grams, ≥ 2500 grams) with three supplemental margins depending on missingness of the variables. For more details about this study, see Baker (1992).

The program in Appendix A.4 coded with S-Plus (version 4.5) is used to fit data according to the nine different models of Baker (1992). In order to avoid recursion problems, we coded model 1 in such a way that the user can determine when the procedure converges by simply calling the function with a fixed number of iterations. When an estimate includes a "such that" statement (see subsection 3.6.2), we compute that estimate by solving a system of linear equations to respect the condition, taking in account the results obtained from the program.

Missing	Smoker	Birth weight (in grams)			
		not missing		missing	
		< 2500	≥ 2500	< 2500	≥ 2500
no	yes	4 511.71	21 007.64	185.75	864.91
	no	3 394.20	24 133.45	139.74	993.60
yes	yes	81.03	215.94	163.68	436.15
	no	60.96	248.06	123.14	501.04

Table 4.8: Estimated expected counts under Model $(\alpha_{.j}, \beta_{..})$

In order to illustrate one of the models, we computed the estimates for model 3. Expected counts (μ_{ijkl} terms) are displayed in Table 4.8. All answers are rounded after the second decimal. Our results perfectly agree with those from Baker (1992). This model has a G^2 value of 0.01 with 1 degree of freedom. Selection of the best model using G^2 and further analysis may be found in Baker (1992).

Chapter 5

Application to the TRICC study

5.1 The TRICC study

The TRICC (Transfusion Requirements In Critical Care) study is a multicenter, randomized controlled trial among critically ill patients with euvolemia after initial treatment who had hemoglobin concentrations of less than 9.0 g per deciliter within 72 hours after admission to the intensive care unit (ICU). Investigators wanted to determine whether a restrictive approach to red-cell transfusion that maintains hemoglobin concentrations between 7.0 and 9.0 g per deciliter is equivalent to a more liberal strategy of maintaining hemoglobin concentrations between 10.0 and 12.0 g per deciliter in the study population of critically ill patients. The primary outcome was death from all causes in the 30 days after randomization. Secondary outcomes included 60-day rates of death from all causes, mortality rates during the stay in the ICU and during hospitalization, and survival times in the first 30 days. Investigators also measured organ failure and dysfunction, including the number and rates of organ failure and the multiple-organ-dysfunction score.

The study includes 22 tertiary-level and 3 community ICU's in Canada and took place from November 1994 to November 1997. The first target sample size was 2300 patients. This enabled investigators to detect an absolute difference of 4 percent in 30-day mortality between the two groups, assuming a combined mortality rate of

18 percent. However, an interim analysis conducted in a blind fashion by the data-monitoring committee after 404 patients revealed that the combined mortality rate was actually 23 percent. This changed the detectable difference. For this reason, the target sample decreased to 1620 patients. This second target sample allowed investigators to detect an absolute difference in the 30-day mortality rate of 5.5 percent between the groups. Overall, 838 patients were enrolled in the study: 418 patients were assigned to a restrictive strategy of transfusion and 420 to a liberal strategy of transfusion. There were no significant differences in any baseline characteristics between the two groups.

The following process has been used to allocate patients. Consecutive critically ill patients with normovolemia were assigned to one of the two treatment groups, stratified according to center and disease severity (an Acute Physiology and Chronic Health Evaluation (APACHE 11) score of 15 or less or a score of more than 15, with higher score indicating more severe disease) and balanced with the use of permitted blocks of four or six. Sealed, opaque envelopes arranged in a computer-generated random order were prepared by the data-coordinating center and distributed to each participating institution, where they were opened sequentially to determine the patients treatment assignments. Other issues such as monitoring, compliance and results may be found in Hébert (1999). All analyses were conducted on an intention-to-treat basis.

5.1.1 Data

Since almost all families presented in this thesis are designed for non-nested patterns of missing data, we needed to create a new data set based on the original data of TRICC. Among the various outcome measures assessed during TRICC study, we focus on the multiple-organ-dysfunction (MOD) score which has a range from 0 to 24, with higher score indicating more severe organ dysfunction. The patient's score was assessed on the day of admission to the ICU.

In order to obtain a variable that includes margins, we created a new variable that looks at the score obtained on renal organ failure and hematologic organ failure, and dichotomized it to 0 if the score was less than 2, and to 1 if the score was 2 or more.

Day 5	Day 10	
	0	1
0	87	47
1	59	90

Table 5.1: Original allocation of patients for the TRICC study

Day 5	Day 10		
	0	1	missing
0	71	33	16
1	49	73	12
missing	5	9	15

Table 5.2: Allocation of patients after randomization for the TRICC study

Then, we kept in the data set all patients who had complete data at days 5 and 10 of the trial, both treatments combined, i.e. 283 patients. This excludes all patients who died or were discharged from the ICU before day 10 (and possibly some who had missing data). Original data are shown in Table 5.1. Finally, we randomly assigned patients to the three margins, with a probability $p = 0.10$ to be missing at the second occasion but not at first; $p = 0.05$ to be missing at the first occasion but not at second; and $p = 0.05$ to be missing at both occasions. The new data set is presented in Table 5.2.

5.1.2 Results

Based on section 3.1, we verify if the naive and unconditional information are equivalent. The model of subsection 3.1.1 is fitted considering only the fully observed data and the row supplemental margin information (as done in Molenberghs (1999a)). Results are shown in Table 22. Standard errors are found using the inverse of the related information matrix. We note that $\hat{\theta}_1$ is quite different from $\hat{\theta}_0$. However, the differences in their standard errors are not significant. Moreover, the difference between

Parameter	Parameter estimate
λ	0.4735 (0.0297, 0.0297)
θ_1	0.5984 (0.0474, 0.0449)
θ_0	0.3173 (0.0427, 0.0410)
η_1	0.9104
η_0	0.8667
η	0.8874

Table 5.3: Parameters (naive standard errors, unconditional standard errors)

$\hat{\eta}_1$ and $\hat{\eta}_0$ is relatively small. From this result, we cannot say without any doubt that the condition for a MCAR mechanism to operate ($\eta = \eta_1 = \eta_0$) is fulfilled, but we can think that a model with a MCAR pattern would give good results.

Results for each family are found in Table 5.4. Note that family I does not use the missing at both occasions cell. This is reflected in the results of Table 5.4. Since family III is designed for nested patterns of missing data, it cannot be applicable to the TRICC study. In family V, the symbol "*" represents a family where boundary estimates of Table 3.8 are used to compute complete data counts.

5.1.3 Analysis

In general, we can say that the results are quite stable across models, either considering complete data counts or odds ratio estimate. Concerning the odds ratio, it varies from 3.20 to 3.27 and because none of the 95% confidence intervals include 1, we could say that it is significant. However, we cannot interpret this odds ratio because of the longitudinal context of this study. Obviously, people who had less than two damaged organs at day 5 have higher chance to still have less than two damaged organ at day 10, than people who had two or more damaged organs at day 5.

The procedure for family I converged quickly with 8 iterations.

Family IIA required 5 iterations to converge. Convergence of families IIB, IIC and IID is relatively slow with 524, 253, and 657 iterations. However, those procedures are not really CPU time consuming, so around 10 seconds are sufficient to get a point

Family	Cell				OR (s.e.)	G^2	d.f.	p
	(1,1)	(1,2)	(2,1)	(2,2)				
I	67.71	32.84	44.47	68.99	3.20 (0.92)	2.91	2	0.2334
IIA	71.50	34.68	46.95	72.85	3.20 (0.89)	2.91	2	0.2334
IIB	72.89	33.88	47.34	71.89	3.27 (0.91)	1.34	1	0.2470
IIC	71.00	33.00	48.44	73.56	3.27 (0.92)	0.11	0	-
IID	74.97	29.19	61.77	77.09	3.21 (0.89)	0.00	0	-
IV	71.50	34.68	46.96	72.85	3.20 (0.89)	2.91	2	0.2334
V1	71.50	34.68	46.96	72.85	3.20 (0.89)	2.91	2	0.2334
V2	69.64	33.71	48.06	74.58	3.21 (0.90)	1.61	1	0.2045
V3	72.89	33.88	47.89	71.34	3.21 (0.90)	1.24	1	0.2655
V4	65.81	31.86	45.42	70.47	3.21 (0.92)	47.42	1	< 0.0001
V5*	77.14	35.86	47.34	71.89	3.27 (0.90)	0.00	1	-
V6*	71.00	33.00	48.44	73.56	3.27 (0.92)	0.00	0	-
V7	71.00	33.00	49.00	73.00	3.21 (0.90)	0.00	0	-
V8*	71.00	33.00	48.44	73.56	3.27 (0.92)	0.00	0	-
V9	71.00	33.00	49.00	73.00	3.21 (0.90)	0.66	0	-

Table 5.4: Model fit (complete data counts; OR (s.e.), odds ratio (standard error); G^2 , likelihood ratio test statistic; d.f., degrees of freedom; p, p-value)

of convergence. The complete data counts for families IIA, IIB, and IIC are similar. However, we do not consider families IIC and IID, since they fit the data perfectly. The fit of family IIA is good, but family IIB has a better fit according to the p-value. Thus, we have some evidence that an MCAR pattern is operating, because under family IIA, the Markov chains for the MOD score process and the nonresponse process are two separate, independent Markov chains.

Family IV converged after 11 iterations. Results for complete data counts are consistent with family IIA.

For family V, under models $(\alpha_i, \beta_{..})$, (α_i, β_i) , and (α_i, β_j) , the equations for an interior ML estimate yielded a negative value for $\alpha_{1.}$. Therefore, we computed boundary estimates for $\alpha_{1.} = 0$ and $\alpha_{2.} = 0$. In all cases G^2 for $\alpha_{1.} = 0$ was larger than for $\alpha_{2.} = 0$. The point of convergence in family V1 is obtained after 7 iterations. We first reject family V4 because of its poor fit of the data, and family V5* because it is

saturated. We do not consider families V6*, V7, V8, and V9 since they fit the data perfectly. Finally, we are left with families V1, V2 and V3. The fit of family V2 is good, but family V3 has a better fit if goodness-of-fit test is the sole criterion. The most parsimonious model that fit the data well is the model $(\alpha_{..}, \beta_{..})$. It corresponds to family V1 and implies the situation where missing in each variable is constant, or in other words, when missing data are MCAR.

Overall, based on the p-values, family V3 presents the best fit. However, families I, IIA, IIB, IV, VI have similar p-values. Thus, we cannot really say which model is the best for the TRICC study. This is probably due to random allocation of missing information in the data set.

A final comment about results of Table 5.4 is concerning the goodness-of-fit statistic, G^2 . According to Fay (1986), a nonzero G^2 is possible even when the count of parameters formally implies zero degrees of freedom (this happened for families IIC and V9). It simply means that the parameter estimates are interior points close to the boundary.

Chapter 6

Conclusion

In this thesis we focused on incomplete categorical data. We have first shown that the common use of the expected information matrix (also known under the name of naive information) is biased even if the missing data pattern is MAR. Furthermore, we have presented how to model missing data mechanism with a 2×2 table of counts with supplemental margins, especially using the TRICC trial.

We consider families I, IIA, IIB, IV, V1, and V3 to be the best candidates to model the TRICC data. Families I, IIA and V1 all correspond to an MCAR pattern of missing data. However, we note that family I leads to different complete data counts, explained by the fact that this family does not use the missing at both occasions cell. Then, we may prefer results of other families that produce equivalent results and use all the information contained in the margins. Nevertheless, families V1 and V3 are the one that requires fewer quantities to be specified. We also verified that family V covers results of almost all other families. Moreover, it is the easiest family to implement computationally because of its capacity to get closed-form estimates.

For the purpose of this thesis, we solely used common model building process to decide whether or not a model fits the data. In a real context, consideration about the plausibility of a model is necessary. Molenberghs (1999a, 1999b) states that one should use as much context-derived information as possible. Prior knowledge can give an idea of which models are more suitable. He also suggests that the use of covariates

helps to explain the dependence between response mechanism and outcomes. For example, he analyzed his data set under two different covariates and compared the results to get a better understanding of the missingness mechanism.

The following are some of the problems related to this research. Contrary to the survey context, nested patterns of nonresponse are more the rule than the exception in biostatistics, particularly in longitudinal studies. For this reason, there is a need to create more models that could support nested patterns of nonresponse. In this thesis, only families I and III present alternative models that fit data that are nested. Also, we note that what we call a "sampling 0", either in one of the margin or in the missing at both occasions cell, could cause problems to all families except family I. For family I, the procedure stops after the first iteration of the EM algorithm, and consequently, a closed-form estimates may be obtained. In other families, estimates are simply impossible to obtain, because that 0-cell implies a division by 0 (or a multiplication by 0, which cancels the result) in the procedure. "Structural 0" in the missing at both occasions cell generates the same difficulty.

Appendix A

S-Plus Programs

A.1 Chen and Fienberg model

```
EM <- function(x,step)
{
  iteration <- function(m,x)
  {
    m2 <- matrix(0,2,2)
    for (i in 1:2)
      for (j in 1:2)
        m2[i,j] <- x[i,j] + x[i,3]*(m[i,j]/(m[i,1]+m[i,2])) + x[3,j]*(m[i,j]/(m[1,j]+m[2,j]))
    return(m2)
  }
  obs <- 0
  R <- 0
  C <- 0
  for (k in 1:2)
    {R <- x[k,3] + R
     C <- x[3,k] + C
     for (l in 1:2)
       obs <- x[k,l] + obs
```

```

    }
N <- obs + R + C
m <- matrix(0,3,3)
for (i in 1:2)
  for (j in 1:2)
    m[i,j] <- (x[i,j]/obs)*N
for (i in 1:step)
  m <- iteration(m,x)
return(m)
}

```

A.2 Stasny's model

```

stasny <- function(model, x, n, step)
{
  modelA <- function(answer,x,n,obs,R,C,M,add1,step)
  {
    recurA <- function(x,n,answer,obs,R,C,M,add1)
    {
      fr <- function(i,pi,pij)
      {
        resr <- 0
        for (j in 1:(n-1))
          {denomi <- 0
            for (k in 1:(n-1))
              denomi <- (pi[k]*pij[k,j]) + denomi
              resr <- ((x[n,j]*pi[i]*pij[i,j])/denomi) + resr
            }
          }
        return(resr)
      }
    }
  }
}

```

```

fj <- function(i,j,pi,pij)
{
  denomi <- 0
  for (k in 1:(n-1))
    denomi <- (pi[k]*pij[k,j]) + denomi
  resj <- (x[n,j]*pi[i]*pij[i,j])/denomi
  return(resj)
}
for (i in 1:(n-1))
  {pi[i] <- answer[i,1]
   epsilon[i] <- answer[i,(n+1)]
   rorr[i] <- answer[i,(n+2)]
   romm[i] <- answer[i,(n+3)]
   for (j in 1:(n-1))
     pij[i,j] <- answer[i,(j+1)]
  }
pi2 <- matrix(0,(n-1),1)
pij2 <- matrix(0,(n-1),(n-1))
xR <- 0
for (i in 1:(n-1))
  {xR <- add1[i]+x[i,n]
   pi2[i] <- (xR + fr(i,pi,pij))/(obs+R+C)
   for (j in 1:(n-1))
     pij2[i,j] <- (x[i,j] + fj(i,j,pi,pij))/(add1[i] + fr(i,pi,pij))
  }
answer <- cbind(pi2,pij2,epsilon,rorr,romm)
return(answer)
}
for (i in 1:step)
  answer <- recurA(x,n,answer,obs,R,C,M,add1)
return(answer)

```

```

}
modelB <- function(answer,x,n,obs,R,C,M,add1,step)
{
recurB <- function(x,n,answer,obs,R,C,M,add1)
{
fr <- function(i, epsilon, pi, pij)
{
resr <- 0
for (j in 1:(n-1))
{denomi <- 0
for (k in 1:(n-1))
denomi <- ((1-epsilon[k])*pi[k]*pij[k,j]) + denomi
resr <- ((x[n,j]*(1-epsilon[i])*pi[i]*pij[i,j])/denomi) + resr
}
return(resr)
}
fv <- function(i,epsilon,pi)
{
denomi <- 0
for (k in 1:(n-1))
denomi <- ((1-epsilon[k])*pi[k]) + denomi
resv <- (M*(1-epsilon[i])*pi[i])/denomi
return(resv)
}
fj <- function(i,j,epsilon,pi,pij)
{
denomi <- 0
for (k in 1:(n-1))
denomi <- ((1-epsilon[k])*pi[k]*pij[k,j]) + denomi
resj <- (x[n,j]*(1-epsilon[i])*pi[i]*pij[i,j])/denomi
return(resj)
}

```

```

}
for (i in 1:(n-1))
  {pi[i] <- answer[i,1]
   epsilon[i] <- answer[i,(n+1)]
   rorr[i] <- answer[i,(n+2)]
   romm[i] <- answer[i,(n+3)]
   for (j in 1:(n-1))
     pij[i,j] <- answer[i,(j+1)]
  }
pi2 <- matrix(0,(n-1),1)
pij2 <- matrix(0,(n-1),(n-1))
epsilon2 <- matrix(0,(n-1),1)
xR <- 0
for (i in 1:(n-1))
  {xR <- add1[i]+x[i,n]
   pi2[i] <- (xR + fr(i,epsilon,pi,pij) + fv(i,epsilon,pi))/(obs+R+C+M)
   epsilon2[i] <- xR/(xR + fr(i,epsilon,pi,pij) + fv(i,epsilon,pi))
   for (j in 1:(n-1))
     pij2[i,j] <- (x[i,j] + fj(i,j,epsilon,pi,pij))/(add1[i] + fr(i,epsilon,pi,pij))
  }
answer <- cbind(pi2,pij2,epsilon2,rorr,romm)
return(answer)
}
for (i in 1:step)
  answer <- recurB(x,n,answer,obs,R,C,M,add1)
return(answer)
}
modelC <- function(answer,x,n,obs,R,C,M,add1,step)
{
recurC <- function(x,n,answer,obs,R,C,M,add1)
{

```

```

fv <- function(i,romm,pi,pij)
{
  resv <- 0
  for (j in 1:(n-1))
    {denomi <- 0
      for (k in 1 :(n-1))
        denomi <- ((1-romm[k])*pi[k]*pij[k,j]) + denomi
      resv <- ((x[n,j]*(1-romm[i])*pi[i]*pij[i,j])/denomi) + resv
    }
  return(resv)
}

fr <- function(i,romm,pi)
{
  denomi <- 0
  for (k in 1:(n-1))
    denomi <- (romm[k]*pi[k]) + denomi
  resr <- (M*romm[i]*pi[i])/denomi
  return(resr)
}

fj <- function(i,j,romm,pi,pij)
{
  denomi <- 0
  for (k in 1:(n-1))
    denomi <- ((1-romm[k])*pi[k]*pij[k,j]) + denomi
  resj <- (x[n,j]*(1-romm[i])*pi[i]*pij[i,j])/denomi
  return(resj)
}

for (i in 1:(n-1))
  {pi[i] <- answer[i,1]
    epsilon[i] <- answer[i,(n+1)]
    rorr[i] <- answer[i,(n+2)]
  }

```

```

    romm[i] <- answer[i,(n+3)]
    for (j in 1:(n-1))
        pij[i,j] <- answer[i,(j+1)]
    }
pi2 <- matrix(0,(n-1),1)
pij2 <- matrix(0,(n-1),(n-1))
romm2 <- matrix(0,(n-1),1)
xR <- 0
for (i in 1:(n-1))
    {xR <- add1[i]+x[i,n]
    pi2[i] <- (xR + fv(i,romm,pi,pij) + fr(i,romm,pi))/(obs+R+C+M)
    romm2[i] <- fr(i,romm,pi)/(fv(i,romm,pi,pij) + fr(i,romm,pi))
    rorr2[i] <- add1[i]/(xR)
    for (j in 1:(n-1))
        pij2[i,j] <- (x[i,j] + fj(i,j,romm,pi,pij))/(add1[i] + fv(i,romm,pi,pij))
    }
answer <- cbind(pi2,pij2,epsilon,rorr2,romm2)
return(answer)
}
for (i in 1:step)
    answer <- recurC(x,n,answer,obs,R,C,M,add1)
return(answer)
}
modelD <- function(answer,x,n,M,N,add1,step)
{
    recurD <- function(x,n,answer,M,N,add1)
    {
        fro <- function(i,pi,pij)
        {
            add10 <- 0
            for (k in 1:(n-1))

```



```

    {add11 <- 0
      for (l in 1: (n-1))
        add11 <- (pi[l]*pij[l,k]) + add11
      add10 <- (x[n,k]*pi[i]*pij[i,k]/add11) + add10
    }
  return(add10)
}
fve <- function(i,pi,pij,romm,add12)
{
  add13 <- 0
  for (k in 1:(n-1))
    add13 <- (M*romm[k]*pi[i]*pij[i,k]/add12) + add13
  return(add13)
}
fbl <- function(i,rorr,pij)
{
  add18 <- 0
  for (k in 1:(n-1))
    add18 <- ((1-rorr[k])*pij[i,k]) + add18
  return(add18)
}
ffu <- function(j,pi,pij)
{
  add14 <- 0
  for (k in 1:(n-1))
    add14 <- (pi[k]*pij[k,j]) + add14
  return(add14)
}
fno <- function(i,rorr,pij)
{
  add16 <- 0

```

```

for (k in 1:(n-1))
  {add17 <- 0
    for (l in 1:(n-1))
      add17 <- ((1-rorr[l])*pij[k,l]) + add17
    add16 <- (x[k,n]*(1-rorr[i])*pij[k,i]/add17) + add16
  }
return(add16)
}
for (i in 1:(n-1))
  {pi[i] <- answer[i,1]
    epsilon[i] <- answer[i,(n+1)]
    rorr[i] <- answer[i,(n+2)]
    romm[i] <- answer[i,(n+3)]
    for (j in 1:(n-1))
      pij[i,j] <- answer[i,(j+1)]
    }
pi2 <- matrix(0,(n-1),1)
pij2 <- matrix(0,(n-1),(n-1))
romm2 <- matrix(0,(n-1),1)
rorr2 <- matrix(0,(n-1),1)
add12 <- 0
add15 <- matrix(0,(n-1),1)
for (j in 1:(n-1))
  for (k in 1:(n-1))
    {add12 <- (romm[k]*pi[j]*pij[j,k]) + add12
      add15[j] <- x[k,j] + add15[j]
    }
for (i in 1:(n-1))
  {pi2[i] <- (add1[i] + x[i,n] + fro(i,pi,pij) + fve(i,pi,pij,romm))/N
    romm2[i] <- 1 - (x[n,i]*add12)/(M*ffu(i,pi,pij))
    rorr2[i] <- add15[i]/(add15[i]+fno(i,rorr,pij))
  }

```

```

      for (j in 1:(n-1))
        pij2[i,j] <- (x[i,j] + (x[i,n]*(1-rorr[j])*pij[i,j]/fbl(i,rorr,pij)) +
          (x[n,j]*pi[i]*pij[i,j]/ffu(j,pi,pij)) +
          (M*romm[j]*pi[i]*pij[i,j]/add12)))/(add1[i]+x[i,n]+fro(i,pi,pij)+fve(i,pi,pij,romm))
      }
    answer <- cbind(pi2,pij2,epsilon,rorr2,romm2)
    return(answer)
  }
  for (i in 1:step)
    answer <- recurD(x,n,answer,M,N,add1)
  return(answer)
}
obs <- 0
R <- 0
C <- 0
for (k in 1:(n-1))
  {R <- x[k,n] + R
   C <- x[n,k] + C
   for (l in 1:(n-1))
     obs <- x[k,l] + obs
  }
M <- x[n,n]
N <- obs + R + C + M
add1 <- matrix (0,(n-1),1)
for (i in 1:(n-1))
  for (j in 1:(n-1))
    add1[i] <- x[i,j] + add1[i]
pi <- matrix (0,(n-1),1)
pij <- matrix (0,(n-1),(n-1))
for (i in 1:(n-1))
  for (j in 1:(n-1))

```

```

    {pi [i] <- add1[i]/obs
      pij [i,j] <- x[i,j]/add1[i]
    }
epsilon <- matrix(0,(n-1),1)
rorr <- matrix(0,(n-1),1)
romm <- matrix(0,(n-1),1)
for (i in 1:(n-1))
  {epsilon[i] i- (obs+R)/(obs+R+C+M)
    rorr[i] <- obs/(obs+R)
    romm[i] <- M/(C+M)
  }
answer <- cbind(pi,pij,epsilon,rorr,romm)
print(model)
if (model == 'A') fanswer <- modelA(answer,x,n,obs,R,C,M,add1,step)
else if (model == 'B') fanswer <- modelB(answer,x,n,obs,R,C,M,add1,step)
else if (model == 'C') fanswer <- modelC(answer,x,n,obs,R,C,M,add1,step)
else fanswer <- modelD(answer,x,n,M,N,add1,step)
rown <- c('1','2','3')
coln <- c('pi','pij_1','pij_2','pij_3','epsilon','rorr','romm')
dimnames(fanswer) <- list(paste(rown), paste(coln))
print(fanswer)
}

```

A.3 Fay's model

```

fay <-function(data,margin1,margin2,iter)
{
  arrf <- array(c(c(0,0,0,0,0,0,0,0),c(0,0,0,0,0,0,0,0)),dim=c(2,2,2,2))
  arrf2 <- array(c(0,0,0,0,0,0,0,0),dim=c(2,2,2))
  arrFtilda <- array(c(c(1,1,1,1,1,1,1,1),c(1,1,1,1,1,1,1,1)),dim=c(2,2,2,2))

```

```

arrFbar <- array(c(0,0,0,0,0,0,0,0),dim=c(2,2,2))
arrFhat <- array(c(c(0,0,0,0,0,0,0,0),c(0,0,0,0,0,0,0,0)),dim=c(2,2,2,2))
for (i in 1:iter)
{
  for (j in 1:2)
    for(k in 1:2)
      {arrf[j,k,1,1] <- data[j,k,1,1]
        arrf[j,k,1,2] <- data[j,k,1,2]*arrFtilde[j,k,1,2]/(arrFtilde[j,1,1,2]+arrFtilde[j,2,1,2])
        arrf[j,k,2,1] <- 0
        arrf[j,k,2,2] <- data[j,k,2,2]*arrFtilde[j,k,2,2]/(arrFtilde[1,1,2,2]+arrFtilde[1,2,2,2]
          +arrFtilde[2,1,2,2]+arrFtilde[2,2,2,2])
      }
  for (j in 1:2)
    for(k in 1:2)
      {arrf2[j,k,1] <- arrf[j,k,1,1] + arrf[j,k,1,2]
        arrf2[j,k,2] <- arrf[j,k,2,1] + arrf[j,k,2,2]
      }
  Fhat <- loglin(arrf,margin1,fit=T)
  Fbar <- loglin(arrf2,margin2,fit=T)
  for(j in 1:2)
    for(k in 1:2)
      for(l in 1:2)
        {arrFbar[j,k,l] <- Fbarfit[j,k,l]
          for(m in 1:2)
            arrFhat[j,k,l,m] <- Fhatfit[j,k,l,m]
          }
  for(j in 1:2)
    for(k in 1:2)
      for(l in 1:2)
        for(m in 1:2)
          arrFtilde[j,k,l,m] <- arrFbar[j,k,l]*arrFhat[j,k,l,m] /(arrFhat[j,k,l,1]

```

```

        +arrFhat[j,k,l,2])
    }
    return(arrFtilde)
}

```

A.4 Baker's model (Family V)

```

baker1 <- function(x,step)
{
  iteration <- function(x,m,obs,R,C)
  {
    m2 <- matrix(0,2,2)
    for (i in 1:2)
      for (j in 1:2)
        m2[i,j] <- ((x[i,j] + x[i,3]*(m[i,j]/(m[i,1]+m[i,2])) + x[3,j]*(m[i,j]/(m[1,j]+m[2,j])))
          *obs)/(obs + R + C)
    return(m2)
  }
  obs <- 0
  R <- 0
  C <- 0
  for (k in 1:2)
    {R <- x[k,3] + R
     C <- x[3,k] + C
     for (l in 1:2)
       obs <- x[k,l] + obs
    }
  mij <- matrix(0,2,2)
  for (i in 1:2)
    for (j in 1:2)

```

```

        mij[i,j] <- x[i,j]
for (i in 1:step)
    mij <- iteration(x,mij,obs,R,C)
alpha <- C/obs
beta <- R/obs
g <- (obs*x[3,3])/(R*C)
answer <- cbind(mij,alpha,beta,g)
rown <- c('1','2')
coln <- c('mi_1','mi_2','alpha','beta','g')
dimnames(answer) <- list(paste(rown), paste(coln))
return(answer)
}

baker2 <- function(x)
{
obs <- 0
R <- 0
C <- 0
for (k in 1:2)
    {R <- x[k,3] + R
     C <- x[3,k] + C
     for (l in 1:2)
         obs <- x[k,l] + obs
    }
mij <- matrix(0,2,2)
mi <- matrix(0,2,1)
beta <- matrix(0,2,1)
alpha <- C/obs
for (i in 1:2)
    {for (j in 1:2)
        mij[i,j] <- (x[i,j]*(x[1,j]+x[2,j]+x[3,j])*obs)/((x[1,j]+x[2,j])*(obs+C))
    }
}

```

```

        mi[i] <- mij[i,1]+mij[i,2]
        beta[i] <- x[i,3]/mi[i]
    }
g <- (obs*x[3,3])/(R*C)
answer <- cbind(mij,alpha,beta,g)
rown <- c('1','2')
coln <- c('mi_1','mi_2','alpha','beta','g')
dimnames(answer) <- list(paste(rown), paste(coln))
return(answer)
}

baker3 <- function(x)
{
obs <- 0
R <- 0
C <- 0
for (k in 1:2)
    {R <- x[k,3] + R
     C <- x[3,k] + C
     for (l in 1:2)
         obs <- x[k,l] + obs
    }
mij <- matrix(0,2,2)
mi <- matrix(0,2,1)
alpha <- matrix(0,2,1)
beta <- R/obs
for (i in 1:2)
    for (j in 1:2)
        mij[i,j] <- (x[i,j]*(x[i,1]+x[i,2]+x[i,3])*obs)/((x[i,1]+x[i,2])*(obs+R))
for (j in 1:2)
    {mi[j] <- mij[1,j]+mij[2,j]

```



```

    alpha[j] <- x[3,j]/mi[j]
  }
g <- (obs*x[3,3])/(R*C)
answer <- cbind(mij,alpha,beta,g)
rown <- c('1','2')
coln <- c('mi_1','mi_2','alpha','beta','g')
dimnames(answer) <- list(paste(rown), paste(coln))
return(answer)
}

baker4 <- function(x)
{
  mij <- matrix(0,2,2)
  mi <- matrix(0,2,1)
  alpha <- matrix(0,2,1)
  g <- 0
  C <- 0
  R <- 0
  noindent for (k in 1:2)    {R <- x[k,3] + R
    for (l in 1:2)
      obs <- x[k,l] + obs for (i in 1:2)
        {C <- x[3,i] + C
          for (j in 1:2)
            mij[i,j] <- (x[i,j]*(x[1,j]+x[2,j]+x[3,j])*obs)/((x[1,j]+x[2,j])*(obs+R))
          }
        }
  for (j in 1:2)
    {mi[j] <- mij[1,j]+mij[2,j]
      alpha[j] <- x[3,j]/mi[j]
    }
  g <- (obs*x[3,3])/(R*C)
  answer <- cbind(mij,alpha,g)
}

```

```

rown <- c('1','2')
coln <- c('mi_1','mi_2','alpha','g')
dimnames(answer) <- list(paste(rown), paste(coln))
return(answer)
}

baker5 <- function(x)
{
  obs <- 0
  R <- 0
  C <- 0
  for (k in 1:2)
    {R <- x[k,3] + R
     C <- x[3,k] + C
     for (l in 1:2)
       obs <- x[k,l] + obs
    }
  mij <- matrix(0,2,2)
  mi <- matrix(0,2,1)
  beta <- R/obs
  for (i in 1:2)
    for (j in 1:2)
      mij[i,j] <- (x[i,j]*(x[i,1]+x[i,2]+x[i,3])*obs)/((x[i,1]+x[i,2])*(obs+R))
  g <- (obs*x[3,3])/(R*C)
  answer <- cbind(mij,beta,g)
  rown <- c('1','2')
  coln <- c('mi_1','mi_2','beta','g')
  dimnames(answer) <- list(paste(rown), paste(coln))
  return(answer)
}

```

```

baker6 <- function(x)
{
  mij <- matrix(0,2,2)
  beta <- matrix(0,2,1)
  C <- 0
  for (i in 1:2)
    {beta[i] <- x[i,3]/(x[i,1]+x[i,2])
      C <- x[3,i] + C
      for (j in 1:2)
        mij[i,j] <- x[i,j]
    }
  g <- x[3,3]/(C*(beta[1]+beta[2]))
  answer <- cbind(mij,beta,g)
  rown <- c('1','2')
  coln <- c('mi.1','mi.2','beta','g')
  dimnames(answer) <- list(paste(rown), paste(coln))
  return(answer)
}

```

```

baker7 <- function(x)
{
  mij <- matrix(0,2,2)
  alpha <- matrix(0,2,1)
  C <- 0
  for (i in 1:2)
    {alpha[i] <- x[3,i]/(x[1,i]+x[2,i])
      C <- x[3,i] + C
      for (j in 1:2)
        mij[i,j] <- x[i,j]
    }
  g <- x[3,3]/(C*(alpha[1]+alpha[2]))

```

```
answer <- cbind(mij,alpha,g)
rown <- c('1','2')
coln <- c('mi_1','mi_2','alpha','g')
dimnames(answer) <- list(paste(rown), paste(coln))
return(answer)
}
```

```
baker8 <- function(x)
{
mij <- matrix(0,2,2)
for (i in 1:2)
  for (j in 1:2)
    mij[i,j] <- x[i,j]
answer <- cbind(mij)
rown <- c('1','2')
coln <- c('mi_1','mi_2')
dimnames(answer) <- list(paste(rown), paste(coln))
return(answer)
}
```

```
baker9 <- function(x)
{
obs <- 0
R <- 0
C <- 0
for (k in 1:2)
  {R <- x[k,3] + R
   C <- x[3,k] + C
   for (l in 1:2)
     obs <- x[k,l] + obs
  }
}
```

```
mij <- matrix(0,2,2)
alpha <- matrix(0,2,1)
beta <- matrix(0,2,1)
for (i in 1:2)
  {for (j in 1:2)
    mij[i,j] <- x[i,j]
    alpha[i] <- x[3,i]/(x[1,i]+x[2,i])
    beta[i] <- x[i,3]/(x[i,1]+x[i,2])
  }
g <- (obs*x[3,3])/(R*C)
answer <- cbind(mij,alpha,beta,g)
rown <- c('1','2')
coln <- c('mi_1','mi_2','alpha','beta','g')
dimnames(answer) <- list(paste(rown), paste(coln))
return(answer)
}
```

Appendix B

Parameter estimation under Stasny's model

Note that the dot notation in a subscript indicates summation over that subscript. The given initial estimates correspond to the closed-form estimates of the parameters appropriate under some of the model.

$$\begin{aligned}x_{(i)}^{(0)} &= \frac{x_{i.}}{x_{..}} \\p_{ij}^{(0)} &= \frac{x_{ij}}{x_{i.}} \\ \xi(i)^{(0)} &= \frac{[x_{..} + R.]}{[x_{..} + R. + C. + M]} \\ \varrho_{RR}^{(0)}(i) &= \frac{x_{..}}{[x_{..} + R.]} \\ \varrho_{MM}^{(0)}(i) &= \frac{M}{[C. + M]}\end{aligned}$$

Model A

Closed-form estimators:

$$\begin{aligned}\hat{\xi} &= \frac{[x_{..} + R.]}{[x_{..} + R. + C. + M]} \\ \hat{\theta}_{RR} &= \frac{x_{..}}{[x_{..} + R.]} \\ \hat{\theta}_{MM} &= \frac{M}{[C. + M]}\end{aligned}$$

Iterative procedure:

$$\begin{aligned}\pi_i^{(v+1)} &= \frac{\left\{ x_{i.} + R_i + \sum_j \left[\frac{C_j \pi_i^{(v)} p_{ij}^{(v)}}{\sum_k \pi_k^{(v)} p_{kj}^{(v)}} \right] \right\}}{\{x_{..} + R. + C.\}} \\ p_{ij}^{(v+1)} &= \frac{\left\{ x_{ij} + \left[\frac{C_j \pi_i^{(v)} p_{ij}^{(v)}}{\sum_k \pi_k^{(v)} p_{kj}^{(v)}} \right] \right\}}{\left\{ x_{ij} + \sum_j \left[\frac{C_j \pi_i^{(v)} p_{ij}^{(v)}}{\sum_k \pi_k^{(v)} p_{kj}^{(v)}} \right] \right\}}\end{aligned}$$

Model B

Closed-form estimators:

$$\begin{aligned}\hat{\theta}_{RR} &= \frac{x_{..}}{[x_{..} + R.]} \\ \hat{\theta}_{MM} &= \frac{M}{[C. + M]}\end{aligned}$$

Iterative procedure:

$$\begin{aligned}\pi_i^{(v+1)} &= \frac{\left\{ x_{i.} + R_i + \sum_j \left[\frac{C_j (1-\xi(i)^{(v)}) \pi_i^{(v)} p_{ij}^{(v)}}{\sum_k (1-\xi(k)^{(v)}) \pi_k^{(v)} p_{kj}^{(v)}} \right] + \frac{M(1-\xi(i)^{(v)}) \pi_i^{(v)}}{\sum_k (1-\xi(k)^{(v)}) \pi_k^{(v)}} \right\}}{\{x_{..} + R. + C. + M\}} \\ p_{ij}^{(v+1)} &= \frac{\left\{ x_{ij} + \left[\frac{C_j (1-\xi(i)^{(v)}) \pi_i^{(v)} p_{ij}^{(v)}}{\sum_k (1-\xi(k)^{(v)}) \pi_k^{(v)} p_{kj}^{(v)}} \right] \right\}}{\left\{ x_{i.} + \sum_j \left[\frac{(1-\xi(i)^{(v)}) C_j \pi_i^{(v)} p_{ij}^{(v)}}{\sum_k (1-\xi(k)^{(v)}) \pi_k^{(v)} p_{kj}^{(v)}} \right] \right\}}\end{aligned}$$

$$\xi(i)^{(v+1)} = \left\{ \frac{x_{i.} + R_i}{x_{i.} + R_i + \sum_j \left[\frac{C_j(1-\xi(i)^{(v)})\pi_i^{(v)}p_{ij}^{(v)}}{\sum_k(1-\xi(k)^{(v)})\pi_k^{(v)}p_{kj}^{(v)}} \right] + \frac{M(1-\xi(i)^{(v)})\pi_i^{(v)}}{\sum_k(1-\xi(k)^{(v)})\pi_k^{(v)}}} \right\}$$

Model C

Closed-form estimators:

$$\begin{aligned} \hat{\xi} &= \frac{[x_{..} + R_{.}]}{[x_{..} + R_{.} + C_{.} + M]} \\ \hat{\varrho}_{RR}(i) &= \frac{x_{i.}}{[x_{i.} + R_i]} \end{aligned}$$

Iterative procedure:

$$\begin{aligned} \pi_i^{(v+1)} &= \frac{\left\{ x_{i.} + R_i + \sum_j \left[\frac{C_j(1-\varrho_{MM}(i)^{(v)})\pi_i^{(v)}p_{ij}^{(v)}}{\sum_k(1-\varrho_{MM}(k)^{(v)})\pi_k^{(v)}p_{kj}^{(v)}} \right] + \frac{M\varrho_{MM}(i)^{(v)}\pi_i^{(v)}}{\sum_k\varrho_{MM}(k)^{(v)}\pi_k^{(v)}} \right\}}{\{x_{..} + R_{.} + C_{.} + M\}} \\ p_{ij}^{(v+1)} &= \frac{\left\{ x_{ij} + \left[\frac{C_j(1-\varrho_{MM}(i)^{(v)})\pi_i^{(v)}p_{ij}^{(v)}}{\sum_k(1-\varrho_{MM}(k)^{(v)})\pi_k^{(v)}p_{kj}^{(v)}} \right] \right\}}{\left\{ x_{i.} + \sum_j \left[\frac{(1-\xi(i)^{(v)})C_j\pi_i^{(v)}p_{ij}^{(v)}}{\sum_k(1-\xi(k)^{(v)})\pi_k^{(v)}p_{kj}^{(v)}} \right] \right\}} \\ \xi(i)^{(v+1)} &= \left\{ \frac{\frac{M\varrho_{MM}(i)^{(v)}\pi_i^{(v)}}{\sum_k\varrho_{MM}(k)^{(v)}\pi_k^{(v)}}}{\sum_j \left[\frac{C_j(1-\varrho_{MM}(i)^{(v)})\pi_i^{(v)}p_{ij}^{(v)}}{\sum_k(1-\varrho_{MM}(k)^{(v)})\pi_k^{(v)}p_{kj}^{(v)}} \right] + \frac{M\varrho_{MM}(i)^{(v)}\pi_i^{(v)}}{\sum_k\varrho_{MM}(k)^{(v)}\pi_k^{(v)}}} \right\} \end{aligned}$$

Model D

Closed-form estimators:

$$\hat{\xi} = \frac{[x_{..} + R_{.}]}{[x_{..} + R_{.} + C_{.} + M]}$$

Iterative procedure:

$$\begin{aligned}
 \pi_i^{(v+1)} &= \frac{\left\{ x_{i.} + R_i + \sum_j \left[\frac{C_j \pi_i^{(v)} p_{ij}^{(v)}}{\sum_k \pi_k^{(v)} p_{kj}^{(v)}} \right] + \sum_j \left[\frac{M \varrho_{MM}^{(v)}(j) \pi_i^{(v)} p_{ij}^{(v)}}{\sum_k \sum_h \varrho_{MM}^{(v)}(h) \pi_k^{(v)} p_{kh}^{(v)}} \right] \right\}}{\{x_{..} + R_{.} + C_{.} + M\}} \\
 p_{ij}^{(v+1)} &= \frac{\left\{ x_{ij} + \left[\frac{R_i (1 - \varrho_{RR}(j)^{(v)}) p_{ij}^{(v)}}{\sum_k (1 - \varrho_{RR}(k)^{(v)}) p_{ik}^{(v)}} \right] + \left[\frac{C_j \pi_i^{(v)} p_{ij}^{(v)}}{\sum_k \pi_k^{(v)} p_{kj}^{(v)}} \right] \right\}}{\left\{ x_{.j} + R_i + \sum_j \left[\frac{C_j \pi_i^{(v)} p_{ij}^{(v)}}{\sum_k \pi_k^{(v)} p_{kj}^{(v)}} \right] + \sum_j \left[\frac{M \varrho_{MM}^{(v)}(j) \pi_i^{(v)} p_{ij}^{(v)}}{\sum_k \sum_h \varrho_{MM}^{(v)}(h) \pi_k^{(v)} p_{kh}^{(v)}} \right] \right\}} \\
 \varrho_{RR}^{(v+1)}(j) &= \frac{x_{.j}}{\left\{ x_{.j} + \sum_i \left[\frac{R_i (1 - \varrho_{RR}(j)^{(v)}) p_{ij}^{(v)}}{\sum_k (1 - \varrho_{RR}(k)^{(v)}) p_{ik}^{(v)}} \right] \right\}} \\
 \varrho_{MM}^{(v+1)}(j) &= 1 - \left\{ \left[\frac{C_j \sum_k \sum_h \varrho_{MM}^{(v)}(h) \pi_k^{(v)} p_{kh}^{(v)}}{\sum_k \pi_k^{(v)} p_{kj}^{(v)}} \right] \right\}
 \end{aligned}$$

Bibliography

- [1] Agresti, A. (1990) *Categorical data analysis*, Wiley, New York.
- [2] Baker, S.G., Rosenberger, W.F. and Dersimonian, R. (1992) 'Closed-form estimates for missing counts in two-ways contingency tables', *Statistics in Medecine*, **11**, pages 643-657.
- [3] Baker, S.G. (1992) 'A Simple Method for Computing the Observed Information Matrix When Using the EM Algorithm With Categorical Data', *Journal of Computational and Graphical Statistics*, **1**, pages 63-76.
- [4] Baker, S.G. (1994) 'Composite linear models for incomplete multinomial data', *Statistics in Medecine*, **13**, pages 609-622.
- [5] Birch, M.W. (1963) 'Maximum Likelihood in Three-Way Contingency Tables', *Journal of Royal Statistical Society, Ser. B*, **25**, p.220-233.
- [6] Chen, T., and Fienberg, S.E. (1974) 'Two-dimensional contingency tables with both completely and partially cross-classified data', *Biometrics*, **30**, pages 629-642.
- [7] Cox, D.R., and Hinkley, D.V. (1974) *Theorical Statistics*, Chapman and Hall, London.
- [8] Deming, W.E., Stephan, F.F. (1940) 'On a least squares adjustment of a sampled frequency table when the expected marginal totals are known', *Annals of Mathematical Statistics*, **11**, pages 427-444.

- [9] Dempster, A.P., Laird, N.M., and Rubin D.B. (1977) 'Maximum likelihood estimation from incomplete data via the EM Algorithm (with discussion)', *Journal or The Royal Statistical Society*, **B39**, pages 1-38.
- [10] Espeland, M.A., and Odoroff, C.L. (1985) 'Log-linear Models for Doubly Sampled Categorical Data Fitted by the EM Algorithm', *Journal of the American Statistical Association*, **80**, 391, pages 663-670.
- [11] Fay, R.E. (1986) 'Causal Models for Patterns of Nonresponse' *Journal of the American Statistical Association*, **81**, 421, pages 354-365.
- [12] Fuchs, C. (1982) 'Maximum Likelihood Estimation and Model Selection in Contingency Tables With Missing Data', *Journal of the American Statistical Association*, **77**, pages 270-278.
- [13] Goodman, L.A. (1970) 'The multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications', *Journal of the American Statistical Association*, textbf65, 329, pages 226-256.
- [14] Goodman, L.A. (1972) 'A General Model for the Analysis of Surveys' *American Journal of Sociology*, **77**, 6, pages 1035-1086.
- [15] Goodman, L.A. (1973) 'Causal Analysis of Data from Panel Studies and Other Kinds of Surveys', *78*, 5, pages 1135-1191.
- [16] Hébert *et al.* (1999) 'A multicenter, randomized, controlled clinical trial of transfusion requirements in critical care', *The New England Journal of Medecine*, **340**, 6, pages 409-417.
- [17] Kenward, M.G., and Molenberghs, G. (1998) 'Likelihood Based Frequentist Inference When Data Are Missing at Random', *Statitital Science*, **13**, 3, pages 236-247.
- [18] Little, R.J.A. (1993) 'Models for Nonresponse in Sample Surveys', *Journal of the American Statistical Association*, **77**, pages 237-250.

- [19] Little, R.J.A, and Rubin, D.B. (1987) *Statistical Analysis with Missing Data*, Wiley, New York.
- [20] Molenberghs, G. *et al.* (1999a) 'Missing data perspectives of the fluvoxamine data set : a review', *Statistics in Medecine*, **18**, pages 2449-2464.
- [21] Molenberghs, G. *et al.* (1999b) 'Nonrandom Missingness in Categorical Data: Strengths and Limitations', *The American Statistician*, **53**, 2, pages 110-118.
- [22] Rubin, D.B. (1976) 'Inference and missing data', *Biometrika*, **3**, pages 581-592.
- [23] Silvey, S.D. (1975) *Statistical Inference*, Chapman and Hall, London.
- [24] Stasny, E.A. (1987) 'Some Markov-Chain Models for Nonresponse in Estimating Gross Labor Force Flows', *Journal of Official Statistics*, **3**, 4, pages 359-373.
- [25] Statistics Canada (1977) *Methodology of the Canadian Labour Force Survey 1976*, Ottawa, Canada.
- [26] Wu, C.F.J. (1983) 'On the convergence properties of the EM algorithm', *Annals of Statistics*, **11**, pages 95-103.