IMPROVEMENT OF PROCEDURES FOR INCOMPLETE CATEGORICAL DATA ANALYSIS

HOO LING PING

UNIVERSITI SAINS MALAYSIA

2008

IMPROVEMENT OF PROCEDURES FOR INCOMPLETE CATEGORICAL DATA ANALYSIS

by

HOO LING PING

**Thesis submitted in fulfillment of the requirements
for the degree of
Doctor of Philosophy**

June 2008

ACKNOWLEDGEMENTS

I would like to thank:

Professor M. Ataharul Islam, my main supervisor, for his overwhelmingly help, support and strong motivation during my research works. This thesis would not be completed without him, and I am extremely grateful for his strong encouragement and patience through very difficult time. ‘Do the best’ is his famous saying that I will never forget.

En. Safian Uda, my co-supervisor, for his advices and recommendation to improve the thesis. Dr Husna Hasan, my representative supervisor, for her guidance and advices to improve the thesis.

Associate Professor Dr. Ahmad Izani bin Md. Ismail, Dean of the School of Mathematical Sciences, for his assistance during my studies especially in hiring me as the part-time tutor when the toughest time of my financial problem. Professor Ong Boon Hua, lecturer of the courses of MAT101 and MAT102, for her patience in guiding and assisting me to handle the tutorial. This has served me more time in my research. Insititut Pengajian Siswazah (IPS) for considering me as the graduate assistant which lightened my financial burden.

My dearest husband, Mr. Leong Kok Huai, for his full care, support, encouragement and patience. My caring parents, Mr. Hoo Ai Swee and Mdm. Ling Chew Guk, my brothers, sisters, sisters-in-law, nephews and nieces for their unbounded encouragement and support.

All the staff members in the School of Mathematical Sciences for their continuing help in completing my thesis. Past and present research fellows in Makmal Siswazah 2 for a friendly environment.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF APPENDIX

LIST OF CONFERENCE PAPERS

1 Methods Analyzing Incomplete Categorical Data, *IMT-GT Regional Conference on Mathematics, Statistics and Applications* organized by USM in Penang, June 2006.

2 EM Algorithm and Analysis of Incomplete Categorical Data, *Seminar Kebangsaan Sains Kuantitatif* organized by UUM in Pulau Langkawi, Dec 2006.

3 Estimation Procedures for Analyzing Incomplete Categorical Data: A Comparison Between Maximum Likelihood Estimation (MLE) and EM Algorithm, *International Conference on Research and Education in Mathematics* organized by UPM in Kuala Lumpur, April 2007.

4 Analyzing Incomplete Categorical Data: A Logistic Regression Approach, *International Medical and Health Congress* organized by USM in Kelantan, May 2007.

5 Linear Models for Analyzing Incomplete Categorical Data, *IMT-GT Regional Conference on Mathematics, Statistics and Applications* organized by USM in Penang, Dec 2007.

PENAMBAHBAIKAN PROSEDUR-PROSEDUR UNTUK ANALISIS DATA BERKATEGORI YANG TIDAK LENGKAP

ABSTRAK

Semasa proses pengumpulan data, kadang-kala kita tidak akan dapat mengumpul semua data yang diperlukan. Ini akan menyebabkan terdapatnya sebahagian data yang tidak lengkap. Kesimpulan yang tidak sesuai akan wujud apabila penyelidik membiarkan, memangkaskan, menapiskan atau menggabungkan data yang tidak lengkap itu. Ini kerana, data itu mungkin mengandungi maklumat yang penting.

Untuk data berkategori yang lengkap, pengganggar kebolehjadian maximum (PKM) dan algoritma penyesuaian berkadaran lelaran (PBL) telah dipertimbangkan untuk mendapat nilai yang dijangka. Bagaimanapun, teknik yang sedia ada untuk menguruskan data berkategori yang tidak lengkap ialah algoritma EM dan PKM.

Objektif utama kajian ini ialah untuk membanding dan memperbaiki algoritma EM, PKM, model linear teritlak (MLT) dengan pendekatan paut gubahan dan regresi lojistik secara penaburan semula bukan sahaja data yang hilang dalam baris ataupun lajur, tetapi juga data yang hilang dalam baris dan lajur bagi jadual kontingensi dua-hala dan tiga-hala yang tidak lengkap. Melalui proses ini, taburan Binomial telah diperiksa sebagai suatu kes khas apabila ia hanya melibatkan jadual kontingensi nx2, teknik Newton-Rapson telah dimasukkan ke dalam PKM untuk mempercepatkan penumpuan dan ujian nisbah kebolehjadian yang melibatkan data yang tidak lengkap untuk pengujian

ketakbersandaran telah diperkenalkan. Dalam kajian ini, mekanisma data hilang telah dipertimbangkan sebagai hilang secara rawak (HSR).

Secara kesimpulannya, kita telah tunjukkan bahawa skema pensampelan untuk Poisson dan Multinomial adalah sesuai diaplikasikan apabila kita mempunyai semua jenis data hilang. Walau bagaimanapun, jika kita mempunyai jadual kontingensi nx2, maka skema pensampelan Binomial boleh dipertimbangkan. PKM telah menunjukkan bahawa ia adalah pilihan terbaik jika dibandingkan dengan algoritma EM. Keadaan ini disebabkan kedua-dua pendekatan itu memberi keputusan yang sama, tetapi PKM memerlukan langkah yang kurang jika dibandingkan dengan algorithma EM. Ini akan menjimatkan masa untuk mendapat keputusan bagi saiz sampel yang besar. PKM dapat memberikan keputusan yang lebih baik jika ia dipertimbang ke dalam kaedah Newton-Raphson. Apabila PKM dimasukkan ke dalam kaedah Newton-Raphson untuk penumpuan, adalah jelas bahawa PKM dan algoritm EM adalah dua jenis algoritma yang berlainan. Didapati juga bahawa regresi logistik boleh digunakan sebagai suatu alternatif untuk memperoleh nilai jangkaan berbanding dengan PKM atau algoritma EM. Kebarangkalian atau ods untuk nilai yang hilang boleh diperolehi dalam sebutan fungsi logit cerapan yang diketahui. Ini dapat membantu kita untuk mendapatkan nilai jangkaan bagi data yang tidak lengkap tanpa menggunakan algoritma tradisi. Pendekatan regresi logistik menunjukkan bahawa ia boleh menganggar nilai pembolehubah kategori secara berkesan apabila kita mempunyai informasi pada pembolehubah kategori yang lain. Oleh sebab itu, kaedah regresi logistik boleh digunakan sebagai kaedah pembezalayan atau klasifikasi dan kaedah ini boleh

diitlakkan untuk sebarang pembolehubah kategori tanpa menjadikan prosedur ini lebih kompleks untuk pengguna umum.

IMPROVEMENT OF PROCEDURES FOR INCOMPLETE CATEGORICAL DATA ANALYSIS

ABSTRACT

During the process of collecting data, sometimes we may not get the fully observed data. This results in partially incomplete data. An inappropriate conclusion may occur when the researchers ignore, truncate, censor or collapse those data as it might contain important information.

For complete categorical data, maximum likelihood estimation (MLE) and iterative proportional fitting (IPF) algorithms have been considered to obtain the expected values. However, the existing techniques to deal with incomplete categorical data are the EM algorithm and the MLE.

The main objective of this study is to compare and improve the EM algorithm, MLE, generalized linear model (GLM) with composite links and logistic regression approaches by redistributing not only for missing row or missing column data, but also for missing row and column data for the two-way and the three-way incomplete contingency tables. Throughout the process, the Binomial distribution has been examined as a special case when it only involves the nx2 contingency table, the Newton-Rapson method has been adopted in the MLE to make a rapid convergence and the likelihood ratio test which involves the incomplete data for testing independence has been introduced. The missing data mechanism is considered as missing at random (MAR) in this work.

As conclusion, we have shown that the Poisson and the Multinomial sampling schemes are suitable when we have all types of missing data.

However, if we have nx2 contingency table, then the Binomial sampling scheme can be considered. The MLE has demonstrated that it is a better choice as compared to the EM algorithm due to the fact that both of these give the same results but the MLE requires less number of steps as compared to the EM algorithm. This will save the time to get the results for large sample size. The MLE can perform better when it is adopted with the Newton-Raphson method. When the MLE is adopted with that of the Newton-Raphson method of convergence, it is clear that the MLE and the EM algorithm are two different kinds of algorithms. It is also revealed that the logistic regression method can be used as an alternative to obtain the expected values as compared to that of the MLE or the EM algorithm. The probability or odds of a missing value in terms of logit function of known observations can be obtained. Without employing traditional algorithm, this helps to get the expected values for incomplete data. The logistic regression approach shows that it can effectively estimate the value of a categorical variable when we have information on the other categorical variables. Hence, the logistic regression method can be used as a discriminant or classification method as well and this method can be generalized for any number of categorical variables without making the procedure more complex for the general users.

CHAPTER 1

INTRODUCTION

1.0 Introduction

The role of data is very important to the statisticians. Without data analysis, statistics will not be complete and illustrative. There are different types of errors in data. This can be human error, equipment error or even unexpected error. For example, some people refuse to fill in the information on a questionnaire or be interviewed due to embarrassing questions like drug abuse, sexual activities, age or income. The other reasons may stem from clinical trials data with some patients not following up their health after certain time period. In engineering, some data may be lost because of mechanical breakdown for an industrial experiment. All the above examples will result in some incomplete data.

The analysis of categorical data rapidly emerged as an important field of research after mid-twentieth century. This is due to the influence of increasing availability of multivariate data sets with categorical responses in the social, behavioural, biomedical sciences, public health, ecology, education, food science, marketing and industrial quality control. Categorical data analysis has provided important insights in resolving problems with categorical response. Since the 1970s, incomplete data analyses have emerged as an important issue of concern. Until today, many methods have become available to analyse

incomplete data. Although the focus is mostly on continuous outcomes, incomplete categorical data have also been well studied. Several developments served special instances of the problems, but the most popular approach has been the EM algorithm. With this background, this study focuses on the importance of the EM algorithm and also on alternative ways to analyse incomplete categorical data.

1.1 Background of Study

An incomplete table is referred to as a table in which the entries are missing, a prior zero or undetermined (Fienberg, 1980). Incomplete data is always the main obstacle for researchers to extend their works. This is especially true for the case of incomplete categorical data. The most common ways for researchers to solve this problem are by ignoring, truncating, censoring or collapsing those data so that their work can be continued. But these are not the wise ways to solve the problems because such procedures may lead to inappropriate conclusion and confusion because those data might contain important information.

Molenberghs and Goetghebeur (1997) defined the missing data mechanism as *ignorable missing data mechanism* and *non-ignorable missing data mechanism.* For ignorable missing data mechanism, it involves the process of missing completely at random (MCAR) and missing at random (MAR). When the missingness is independent of both unobserved and observed data, the

non-response process is named as MCAR. However, if conditionally on the observed data, the missingness is independent of the unobserved measurements, the non-response process is called as MAR. The informative process is for non-ignorable or informative missing data mechanism. In other words, the process is termed as *informative* when the process is neither MCAR nor MAR. The *informative* mechanism is based on some information regarding the missing data regarding the pattern of missing data, unlike noninformative mechanism such as MCAR and MAR

The problem of estimation for incomplete contingency table under the quasi-independence model was examined by Fienberg (1970). Fienberg used the maximum likelihood estimation (MLE) procedure. Similarly, the MLE for the Poisson and the Multinomial sampling distributions for the incomplete contingency tables in the presence of missing row and missing column data were considered by Chen and Fienberg (1974). Chen and Fienberg (1976) extended their works which focused on cross-classifications containing some totally mixed up cell frequencies with the Multinomial sampling. In the following year, Dempster, Laird and Rubin (DLR) (1977) presented the MLE of incomplete data and named the algorithm as the EM algorithm since each iteration of the algorithm involves expectation (E) and maximization (M) steps. This method has been used extensively by other researchers especially for incomplete categorical data. Among many others, Fuchs (1982), Nordheim (1984), Fay (1986), Baker and Laird (1988), Philips (1993) have used the EM algorithm for analyzing

incomplete categorical data. Baker (1994) and Galecki, Have and Molenberghs (2001) incorporated the Newton-Raphson approach into the EM algorithm to improve the convergence of the EM. The EM algorithm is well developed (Lauritzen, 1995) to exploit the computational scheme of Lauritzen and Spiegelhalter (1988) to perform the E-step of the EM algorithm to find the MLEs in hierarchical log-linear models and recursive models for contingency tables with missing data. Besides, Molenberghs and Goetghebeur (1997) presented a simple expression of the observed data log-likelihood for the EM algorithm.

Since the EM algorithm has been introduced, the MLE procedure is ignored by the researchers until 1985. Then Stasny (1985) used the MLE to process the model based on data from Current Population Survey and the Labour Force Survey to estimate the gross flow data. Most recently, Lyles and Allen (2003) proposed the MLE procedure with the Multinomial likelihood properly accounting for missing data and assumed that the probability of missing exposure depends on true exposure.

In another development, Rindskopf (1992) has considered the generalized linear models with composite links to fill in contingency tables with supplementary margin, and which fits loglinear models when data are missing.

Little and Schluchter (1985) have considered the logistic regression and the discriminant analysis with missing predictors and unclassified observations to obtain the maximum likelihood estimation for mixed continuous and

categorical data with missing values. Vach and Schumacher (1993) have compared the approaches among maximum likelihood estimation, pseudo maximum likelihood estimation and probability imputation for the logistic regression analysis with incompletely observed categorical covariates. However, Fitzmaurice *et al*. (1996a) considered the logistic regression as a likelihood-based regression model to analyze binary data with attrition and in the same year, Fitzmaurice *et al.* (1996b) have considered the same method to model the association between the binary response in terms of conditional log odds ratios. Besides, the logistic regressions have also been considered by Ibrahim *et al.* (1999) to propose estimating parameter in the generalized linear models with missing covariates. They considered conditional distribution consisting of logistic regression. James (2002) extended the generalized linear models to the situation where some of the predictor variables are observations from a curve or function. He considered this approach to perform linear, logistic and censored regression with functional predictor in missing data problems.

1.2 Objectives of Study

The specific objectives of the study are listed below:

1. To show that the MLE and the EM algorithm can be extended to estimate when row and column data are missing based on various assumptions and by considering the different types of distributions for the two-way and the three-way incomplete contingency tables.

2. To show the Binomial sampling scheme for the MLE and the EM algorithm as a special case of the Multinomial sampling scheme.
3. To show that the likelihood ratio test procedure can be used for the incomplete categorical data based on the MLE and the EM algorithm.
4. To show that the MLE procedure can be improved by using the Newton-Raphson method for quick convergence in case of missing categorical data.
5. To compare the revised the MLE and the EM algorithm.
6. To show how the GLM with composite links can be employed to deal with the missing row and column data in the two-way and the three-way contingency tables.
7. To compare among the GLM procedure with the MLE and the EM algorithm.
8. To show that the proposed method, the logistic regression procedures can be used to estimate the missing categorical data for three variables which can be extended for more variables.
9. To compare the proposed method with existing methods such as the EM algorithm employing various test procedures such as apparent error rate (APER), sensitivity, specificity, McNemar test and Mantel-Haenszel test.

1.3 Organization of Thesis

Organization of the thesis is as follows: Chapter 2 is literature review of the study. Chapter 3 is a review of existing methods for complete and incomplete contingency tables. This includes two-way and three-way contingency tables. The maximum likelihood estimation (MLE) and the iterative proportional fitting (IPF) methods are applied to the log-linear models to obtain the expected values for complete contingency tables. However, for incomplete contingency tables, the EM algorithm and the MLE have been reviewed. Then likelihood ratio test is considered for testing independence for both complete and incomplete contingency tables.

Chapter 4 contains the method to estimate the missing value which is the EM algorithm. In this chapter, the Multinomial and the Poisson sampling to redistribute not only missing row and missing column data but also missing row and column data are been considered. However, the Binomial sampling has been demonstrated that it can be considered as the special case of redistributing missing column data when we have nx2 contingency tables. Besides, incomplete three-way contingency table for all types of possible missing data in the Poisson and the Multinomial sampling also be considered. Likelihood ratio test for the test of independence of the expected value for two-way and three-way contingency tables are considered.

In Chapter 5, estimation and test procedures are considered for the linear models. Loglinear model is considered for incomplete data and the MLE procedure is demonstrated and then the generalized linear model (GLM) with composite links are also shown. In this chapter, not only missing row, missing column data but also missing row and column data on the Poisson and the Multinomial sampling are distributed for the MLE while the Binomial sampling as a special case to redistribute missing column data when we have nx2 contingency table for the MLE is considered. The results of the MLE are shown with an application and a test for the independence is highlighted on the basis of the likelihood ratio test. Then, the MLE is compared with the EM algorithm. The MLE is improved by adopting the Newton-Raphson method. However, for the GLM with composite links, all possible types of missing data in two-way and three-way contingency tables have been considered. Results of the GLM with composite links are compared with the MLE and the EM algorithm. Then the weaknesses of the GLM with composite links are highlighted also in this chapter by comparing it with the MLE and the EM algorithm.

The conditional distribution by using the logit link functions has been considered and this method will be discussed in Chapter 6. The logistic regression model is employed to estimate the incomplete categorical data for three-way contingency tables where one of the variables is let to be the outcome variable and the other two as independent variables. After that, this result has been compared with that of the EM algorithm. The suitability of the method is

examined on the basis of the apparent error rate, sensitivity, specificity, McNemar's test and Mantel-Haenszel test. Chapter 7 includes the overall conclusion of the study and comments on future research.

CHAPTER 2

LITERATURE REVIEW

2.0 Introduction

One of the main obstacles for researchers is to deal with incomplete data in studies related to categorical data. In this chapter, a review of the literature is provided in this context. It is observed that the EM algorithm, the MLE, and the logistic regression models have been employed by various researchers in the field of incomplete categorical data. This chapter includes a review of the important studies, in relation to the analysis of incomplete categorical data.

2.1 EM Algorithm

In 1977, a broadly applicable algorithm for computing the maximum likelihood estimates from incomplete data is proposed by Dempster, Laird and Rubin. They proposed an algorithm which is named as EM algorithm because it involves expectation step (E-step) and maximization step (M-step) in each iteration. They have derived the monotone behavior of the likelihood and convergence of the algorithm.

It is evident that categorical data are often collected with some incomplete data records (Fuchs, 1982). There are two general methods to categorize the incomplete data: (i) data are summarized in a single table with missing or a priori empty or combined categories, or (ii) data summarized in the form of two or more related tables, one fully categorized and the other containing data that are only partially categorized. Fuchs (1982) states that single tables with missing or a priori empty cells can be analyzed by fitting the

log-linear models (Bishop, Fienberg, and Holland, 1975; Haberman 1974b, 1979). Tables with combined categories can be caused by truncated or censored data (Hartley, 1958) or by contingency tables with mixed-up cells (Haberman, 1974a; Chen and Fienberg, 1976).

Data summarized into a series of mutually exclusive tables, only one of which is fully categorized, may arise when the investigator collects the data by refining categories for a subsample and by grosser categories for the remainder of the sample. Analysis of such partially categorized data is considered by Hocking and Oxspring (1971, 1974) and by Chen and Fienberg (1974). A special case of such a series of tables occurs when data are missing for one or more of the categories of variables. The partially categorized tables contain those observations that cannot be included in the fully categorized table and each observation is included in the highest-order table to which it can be assigned. Chen and Fienberg (1974) illustrated the expected frequencies in the main table and the process by which some observations lose their row or column identity for two-way contingency tables with supplemental row and column subtables.

Fuchs (1982) considered data classified into a multiway frequency table where the values of all the variables are recorded for a subset of the sample, while other subsamples have data missing for one or more variables. Since missing observations can occur on any variables, many of the supplemental tables may be sparse.

Fuchs (1982) applied the EM algorithm (Dempster, Laird and Rubin, 1977) for the problem he had considered to obtain the maximum likelihood

estimates (MLEs) for the expected cell frequencies in tables augmented by incomplete data. The factorization of the likelihood (Rubin, 1974) is applied for data with a nested pattern to obtain MLEs for the saturated model. Tests of fit for the log-linear models in the presence of incomplete data are also considered. It is shown that the allocation of the incomplete data according to a specific model may affect the tests of fit considerably. Therefore, fitting the model and computing the MLEs are suggested in two separate stages. For the purpose of application, Fuchs (1982) considered the data from the extensively argued Protective Services Project for Older Persons (Blenkner, Bloom, and Weber, 1974). The data on all the variables were available for 101 participants. The data not available were physical status with 1 frequency, mental status with 33 frequencies and physical status and mental status with 29 frequencies. The data is presented in the Appendix, Table A.1

It is evident from the conclusion of Fuchs (1982) that the algorithms required for computing the MLEs in frequency tables formed from an incomplete data matrix are not much more difficult than those in the case of complete data. The algorithms used in the case of complete data can be either used iteratively to yield the MLEs for the incomplete data case or modified to yield the MLE in a single cycle. Fuchs (1982) also found that the increased reliability of the results, the ease of computation, and the intuitive interpretation are appealing features of the procedure.

It was observed by Fay (1986) that a nonresponse originated from a questionnaire might pose an obstacle to get the complete survey data. Fay (1986) states that the most frequently cited reasons for nonresponse in sample

surveys is attributable to unwillingness of respondents to provide the correct information. However, besides the cause of human behavior, it might also be caused by inability of respondents to understand a question or lack of knowledge of the respondents. Another important reason of nonresponse is due to propagation of erroneous nonresponse as a sequel to a single question. The study of Fay (1986) was limited to nonresponse for categorical data. Therefore a general class of models to process the nonresponse to represent a different orientation to the problem of inference from the observed data is presented. The data considered by Fay (1986) is on survival of subjects cross-classified by initial evaluations of physical and mental status. This data set is shown at Appendix, Table A.2.

According to Fay (1986), many possible models are available for three or more variables. The causal models discussed about forming of a rich class of alternatives. There has been utmost attention in the literature on the model-based approaches for analysis of missing data in the case of given ignorability of response and it occupies a central role but the assumption of ignorability may not always be correct. Additional conditions are required for causal models which may cause substantially larger effect on the overall variance. This becomes the disadvantage of causal models. As stated by Rubin (1978), the importance of appreciating two sources of uncertainty in the analysis of data subjects to nonresponse are: i) imputation does not properly include the random contribution or variance due to treatment of imputed data; and ii) the source of uncertainty in the analysis can be stemmed from selection of a model or assumption for the process of nonresponse. Therefore causal models may serve as an alternative for deciding the effect of the choice of model.

As referred to Baker and Laird (1988), a common problem in the analysis of survey data involves incomplete data with a possible nonignorable response mechanism. The response mechanism (the reason whether or not a unit response is obtained) is said to be nonignorable if it depends on a subject's unobserved response (Little, 1982). The example considered by Baker and Laird (1988) is the use of polling data to predict the proportion of voters preferring Truman won with 52% of the two party votes, although polling results predicted Dewey. The set of data is shown at Appendix, Table A.3. The causes for failure of the polls in an election were studied extensively by a special committee of the Social Science Research Council (SSRC) and discussed in its report (Mosteller et al., 1949). This report identified several weaknesses of the polls, including failure to weight sample data appropriately, heavy vote switching in the two weeks prior to the election, and nonresponse bias. The purpose of Baker and Laird (1988) was to illustrate that an appropriate model could be used effectively to adjust for nonresponse bias.

Baker and Laird (1988) modeled the response mechanism associated with a categorical outcome and a set of covariates by using the log-linear models. The results of estimation and hypothesis testing can be sensitive to the choice of model for the response mechanism. By using different regression models, the model for response mechanism can be systematically varied and the sensitivity of estimates can be checked and tested to a variety of plausible assumptions for nonresponse.

It is evident from the conclusion and finding of Baker and Laird (1988) that their results may be sensitive to the model selected. Responses were

obtained on a subset of nonrespondents through intensive pursuit of a random fit of a richer set of alternative models for analyzing categorical data subject to nonresponse. In the M step of the EM algorithm, the log-linear models also make it easy to examine all plausible models for the response mechanism, which is necessary for gauging the uncertainty due to nonresponse in estimation and hypothesis testing. For incomplete 2x2x2 table, it can be demonstrated that some models are not estimable, solutions may occur on the boundary, and G^2, goodness of fit may be nonzero even when degree of freedom (df) (lack of fit) = 0. It is difficult to determine which models are estimable and to count df (lack of fit) with large tables. As a result, model comparisons involving ΔG^2 and Δdf should be done carefully to ensure that df is counted correctly.

Ibrahim (1990) examined the general problem of incomplete data for any generalized linear model (GLM) with discrete covariates, in which incompleteness is due to partially missing covariates on some observations. The EM algorithm is applied to obtain the MLEs. Under some very general conditions, the E-step of the EM algorithm can be written as a weighted complete data log likelihood for any GLM is shown.

The example which Ibrahim (1990) considered to illustrate the method of weights for the EM algorithm involves a logistic regression from a data set of incomplete observation. The example involves a study of 82 patients who experienced translaryngeal intubation (TLI) for more than four days and were prospectively evaluated for laryngeal complications. The purpose of the study was to identify a group of patients experiencing prolonged TLI (more than four

days) and to prospectively evaluate the incidence and type of laryngeal complications they might suffer. Data were collected on the patients regarding 13 baseline explanatory variables (covariates) during the period of TLI. From 13 covariates, 3 are continuous and 10 are dichotomous and among that 13 variables, 4 are incomplete. For these data, the response variable, y, is dichotomized as 0, for no damage and 1 for damage of the larynx at baseline respectively. The three covariates are serum albumin, x_1, which is dichotomized and takes the values of 0 if <30 SI and 1 if ≥30, where SI denotes standard international units. Serum creatinine, x_2, is second explanatory variable. This is also dichotomous and takes the value 0 if <200 SI and 1 if ≥200 SI. The third covariate, x_3, is the ratio of laryngeal size to tracheal tube size, which is also dichotomized and takes the value 0 if the ratio is less than 0.45 and 1 if the ratio is greater than or equal to 0.45. This data set is illustrated in Appendix, Table A.4.

It is evident from the conclusion of Ibrahim (1990) that the direct use of the Newton-Raphson method on the incomplete data likelihood can also be used to estimate the parameters in an incomplete data problem. For the class of GLM, finding the incomplete data likelihood directly is generally quite difficult, and in most cases it cannot be expressed in a reasonable closed form. Thus carrying out the direct Newton-Raphson on the incomplete data likelihood in GLM is not practical for most situations. The EM algorithm by the method of weights does not require the computation of the incomplete data likelihood. Its entire computation depends only on complete data quantities. The EM algorithm seems to be a more practical approach than the direct Newton-Raphson for the class of GLM. The EM algorithm by the method of weights is not restricted only

to the class of GLM. The idea is actually very general and may be applied to other types of models, such as nonlinear regression models or time series models. Slow convergence rate is the drawback of the EM algorithm.

The analysis of categorical data will not be clear when there are partially classified observations with missing values on one variable (Phillips, 1993). Ignoring partially classified observations do not affect any comparisons of models made by using the likelihood ratio goodness-of-fit statistics if the observations are 'missing at random' (MAR) (Rubin, 1976).

Phillips (1993) has concentrated on three-dimensional tables with missing values on one variable. General results given by Fuchs (1982) and Nordheim (1984) were applied and extended by Phillips (1993). The case of two dimensional tables has been tackled by Little and Rubin (1987). The data set considered by Phillips (1993) is a clinical trial conducted at two centers to compare two drugs. This data is shown at Appendix, Table A.5.

It is evident from the conclusion of Phillips (1993) that when the effect on inferences of using the likelihood ratio test statistic with the assumption that 'missing at random' is relaxed can only be applied on IxJx2 three-dimensional tables. It is possible to obtain the maximum likelihood estimator for the expected cells.

A general-likelihood-based theory for the analysis of data obtained from sample surveys of finite populations have been presented by Breckling et al. (1990). In contrast with the EM algorithm, this method produces explicit expressions for the score and information functions generated by the observed

data which allow us to compute approximate standard errors and test statistics based on these functions. Chambers and Welsh (1993) applied the methods of Breckling *et al.* (1990) to fit the log-linear models to categorical survey data which are subject to non-response. Chamber and Welsh (1993) have made a double contribution, which is a very general class of the log-linear models an algorithm for fitting these models, which yields explicit estimates, are used to quantify the sensitivity of the inferences to the model for non-response.

Little (1985), Fay (1986), Little and Rubin (1987), and Baker and Laird (1988), Chamber and Welsh (1993) tackled the problem of uncertainty about the nature of the non-response mechanism by explicitly considering several competing plausible models for the mechanism. Explicit allowance for simultaneous adjustment for sample design effects is made and explicit expressions for the score and information functions generated by the observed data are produced by Chamber and Welsh (1993). The log-linear models, which are used by Chamber and Welsh (1993), are most closely related to Baker and Laird (1988) but they do not allow for either non-nested patterns of non-response or simultaneous adjustment for sample design effects. Little (1985) and Fay (1986) allow for general (non-nested) patterns of non-response but not for simultaneous adjustment for sample design effect and do not use the log-linear models. Fay (1986) also develops the use of causal models for non-response. The explicit formula of Chamber and Welsh (1993) avoid the need to use the jackknife (Fay, 1986) and bootstrap (Baker and Laird, 1988) to obtain approximate standard errors. Table A.6, in Appendix was used by Chamber and Welsh (1993) to illustrate their methods.

Chambers and Welsh (1993) have demonstrated the construction, interpretation and fitting of the computable log-linear models to categorical survey data with nonignorable nonresponse. This is both feasible and straightforward. They referred to the works of Little (1985), Fay (1986), Little and Rubin (1987), and Baker and Laird (1988) and modeled conditional nonresponse probabilities on the basis of the classification given separately employing the classification probabilities. Their general model formulation for the nonresponse model allows this model to depend on scores, discrete covariates, continuous covariates or a mixture of types of covariates. They obtained explicit expressions for the score and information functions generated by the observed data, which allow computing approximate standard error and test statistics based on these functions. Unfortunately, Chambers and Welsh (1993) cannot guarantee the accuracy of the inferences when the underlying nonresponse is in fact nonignorable. Therefore it is quite difficult to determine from the survey data alone when the nonresponse is ignorable. However, large sudden changes in fitted values as fitted nonignorable nonresponse models become more extreme which may indicate the presence of ignorable nonrepsonse.

Lauritzen (1995) has shown that the computational scheme of Lauritzen and Spiegelhalter (1988) to perform the E-step of the EM algorithm when applied to finding the maximum likelihood estimates or the penalized maximum likelihood estimates in the hierarchical log-linear models and recursive models for contingency tables with missing data can be exploited.

It is evident from the conclusions of Lauritzen (1995) that the procedure of Lauritzen and Spiegelhalter (1988) is able to calculate the term in the E-step of the EM algorithm for the hierarchical log-linear models and recursive models. When the results tend to converge on the basis of the recursive models, the EM algorithm converges at a relatively slower pace. Therefore the relative computational scheme could be used for the probability propagation in the E-step as an alternative which described by Shenoy and Shafer (1990).

According to Molenberghs and Goetghebeur (1997), the most popular approach to analyse incomplete data has been the EM algorithm (Dempster et al., 1997). Definition of complete data is containing observed and unobserved outcomes and full data is combined by set of complete data and missingness indicators (Molenberghs and Goetghebeur, 1997).

From the terminology of Little and Rubin (1987), if the missingness is independent for both observed and unobserved data, then the non-response process is said to be missing completely at random (MCAR). However, if the missingness is independent of the unobserved measurement conditionally on the observed data, then the non-response process is called as missing at random (MAR). If it is neither completely random nor random, then it is termed as *informative*. From previous work of Baker and Laird (1988), Chambers and Welsh (1993) found that the EM algorithm is general, stable and can be implemented conveniently. However, slow rate of convergence and lack of the direct provision of a measure of precision for the estimators have made the EM algorithm not a perfect approach. Therefore Molenberghs and Goetghebeur (1997) proposed an alternative general approach which has advantages when

fitting models to a broad class of incomplete categorical data. A simple expression for the observed data likelihood and its derivatives in terms of the complete data model, provided that the observed data are linear functions of the complete data was presented.

It is evident from Molenberghs and Goetghebeur (1997) that no greater complexity arises when constructed and fitted the observed data likelihood directly rather than for the complete data. Faster convergence and variance estimator at each step of iteration is obtained by considering Fisher scoring algorithm to find the maximum.

As referred by Galecki *et al.* (2001), the advantages for the EM algorithm are its generalizability and stability. But the EM algorithm performs at a slow rate of convergence and lacks straightforward estimation of the precision of parameter estimates. Baker (1994) tried to improve the EM algorithm by incorporating the Newton-Raphson approach. However, Molenberghs and Goetghebeur (1997) proposed a method using the Fisher scoring to maximize the observed likelihood instead of complete data likelihood under a multivariate generalized logistic model with composite link function (McCullagh and Nelder, 1989; Lang and Agresti, 1994; Balagtas *et al.*, 1995; Yang and Becker, 1997). Molenberghs and Goetghebeur (1997) have shown faster convergence and easily yielded variance estimates as part of the Fisher scoring algorithm.

It is evident from the conclusion that Galecki *et al.* (2001) is able to propose a more flexible alternative. The differences between Galecki *et al.* (2001) with Molenberghs and Goetghebeur (1997) are Galecki *et al.* (2001) proposed a more flexible inversion technique for obtaining cell probabilities by

considering the multivariate generalized linear models. Galecki *et al.* (2001) applied an extension of the iterative proportional fitting (IPF) to the inversion process. The IPF approach is used to obtain the maximum likelihood estimates under a hybrid marginal log-linear model. Secondly, the difference between Galecki et al. (2001) with Molenberghs and Goetghebeur (1997) is that they considered iteratively reweighted least squares (IRLs) techniques rather than the Newton-Raphson approach.

According to Tang *et al.* (2007), the EM algorithm is the most widely used approach for finding the maximum likelihood estimate for incomplete-data problems but it lacks the direct provision of a measure of precision for the estimators and the slow rate of convergence. Louis (1982) suggested obtaining the asymptotic variance-covariance matrix of the MLE. The delta method (Tanner, 1996) can be utilized to calculate standard errors of various functions of the cell probabilities.

Tang *et al.* (2007) proposed a novel data augmentation (DA) scheme which involves fewer latent variables and results in a new and efficient EM algorithm. Two bootstrap confidence intervals (CIs) are recommended for small-sample data, for functions of cell probabilities via the new EM algorithm.

Tang *et al.* (2007) reveals that the method they suggested converged much faster than the EM algorithm based on the conventional DA scheme. Also by comparing with the delta methods, the proposed bootstrap methods are feasible and perform well. Throughout the work of Tang *et al.* (2007), they considered the mechanism of MAR.

2.2 Maximum Likelihood Estimation (MLE)

It is evident that two types of empty cells we have to encounter in categorical data analysis are *sampling zero* and *a priori zero* (Fienberg, 1970). Sampling zero occurs due to sampling variability and the relative smallness of the cell probability. Sampling zeros will be eliminated when the sample size is increased. However, a priori zero occurs when the observations are missing or truncated (Goodman, 1968; Watson, 1956). Besides, impossible observation counts is also defined as a priori zero (Bishop and Fienberg, 1969; Mantel and Halperin, 1963; Pearson, 1930, Waite, 1915).

It was observed by Fienberg (1970), most of the authors tried to solve the problem of a priori zero cell by considering a multiplicative model for the non-zero expected cell frequencies. Goodman (1968) introduced the term 'quasi-independence' to describe that multiplicative model. Various procedures for calculating estimates of the expected cell frequencies based on the assumption that the unique maximum likelihood estimates for the quasi-independence model do exist have been proposed by Bishop and Fienberg (1969), Caussinus (1965), and Goodman (1964, 1968). By adopting these procedures, Fienberg and Holland (1970) proposed a different approach. Unfortunately, the existence of unique maximum likelihood estimates for the expected cell counts under the various models could not be successfully shown. Therefore, Fienberg (1970) tried to work on the problem discussed above.

In his study, Fienberg (1970) was able to provide two conditions to prove the existence of unique non-zero maximum likelihood estimates for an incomplete two-way table. Two conditions provided by Fienberg (1970) were: (i)

the marginal totals of row and column are all positive, and (ii) the observed table of cell counts corresponding to the subtable is inseparable.

According to Chen and Fienberg (1974), an unrestricted estimation of the multinomial cell probabilities has been considered by Blumenthal (1968) for some partially cross-classified contingency tables. Hocking and Oxspring (1971) have considered the same problem and they considered the original multinomial corresponding to a two-dimensional cross-classification, and Reinfurt (1970) considered that there are two supplemental multinomial samples corresponding to the row totals and column totals. Blumenthal (1968) noted that results for the random partial classification problem can be essentially the same as the supplemental-sample problem. However, Koch and Reinfurt (1970) and Koch *et al.* (1972) use a modified minimum chi-squared approach to various contingency tables cases of the Hocking-Oxspring (1971) problem. Many of these earlier results are applicable to problems involving the unrestricted estimation of cross-classified cell probabilities. But unfortunately none of the researchers deal with reduced parameterizations for general interest like independence of variables corresponding to rows and columns.

Chen and Fienberg (1974) have shown that their results can be specialized to yield the estimators and asymptotic variances given in Blumnethal (1968). Besides, methods for obtaining the maximum likelihood estimates for expected cell values in contingency tables with partially cross-classified data were described. The log-linear model was considered to deal with the multi-dimensional contingency tables with some partially cross-