Some Topics in Missing Data and Adaptive Confidence Intervals

by

Yan Zhou

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2009

Doctoral Committee:

Professor John D. Kalbfleisch, Co-Chair
Professor Roderick J. Little, Co-Chair
Associate Professor Moulinath Banerjee
Associate Professor Michael R. Elliott

To my baby, my husband and my parents

ACKNOWLEDGEMENTS

I would like to express my appreciation to those people who helped me with my doctoral study and my dissertation. First of all, I would like to express my deep thanks to Dr. John D. Kalbfleisch and Dr. Roderick J. Little, my advisors, for their insightful guidance, encouragement, patience and support in the accomplishment of the whole dissertation. I am very grateful to Dr. Michael R. Elliott and Dr. Moulinath Banerjee, my dissertation committee members, for their time and efforts. Personally, I would like to thank my husband Yuezhou Jing, who has always been devoted and supportive. His love and encouragement helped me through every step of my doctoral study. I also want to thank my baby Balwyn Jing, who makes me understand more meanings of life and gives me endless happiness. I am especially grateful to my father Nianli Zhou and my mother Guichun Wang, for sacrificing so much for my education; without them I would never have the opportunity to complete the doctoral study.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

.

ABSTRACT

Some Topics in Missing Data and Adaptive Confidence Intervals

by
Yan Zhou

Co-Chairs: John D. Kalbfleisch and **Roderick J. Little**

When data are missing at random, the missing-data mechanism can be ignored but this assumption is not always intuitive for general patterns of missing data. In part I, we consider maximum likelihood (ML) estimation for a non-ignorable mechanism which is called almost missing at random (AMAR). We examine in some detail the case of two multinomially distributed categorical variables X and Y, for which X is missing completely at random and Y is MAR given the value and missingness of X. In this case, although ML can be fitted using the EM algorithm, we find non-iterative ML estimates sometimes exist, with some data being excluded for estimating the parameters of interest. A variation of this type of mechanism is also discussed. We apply the AMAR models to data from the Muscatine Coronary Risk Factor Study (Woolson and Clark, 1984).

In part II, we consider one extension of AMAR. Besides two variables with missing values, there is an additional fully observed covariate. Specifically, we consider randomized clinical trials when there is non-compliance with the assigned treatment and subsequent non-response. We build a connection between AMAR and latent ignorability (Frangakis and Rubin, 1999). To identify the model, we further specify two assumptions for principal compliance and two assumptions for missingness of the outcome. In each of four scenarios defined by combinations of these assumptions, we derive ML estimates by using the EM algorithm, as well as non-iterative ML estimates by implementing pattern-mixture models with covariates (Little and Wang, 1996). The later approach shows that, under certain conditions, the method-of-moments estimates are also ML estimates. We show that the models for principal compliance determine which type of analysis is used to estimate treatment efficacy, per-protocol analysis or IV estimation with the treatment assignment indicator as the instrumental variable. On the other hand, we show that the assumptions for missing outcome determine whether non-iterative ML estimates exist or not. We apply our methods to data from a double-blinded randomized clinical trial with clozapine vs. haloperidol for patients with refractory schizophrenia. (Rosenheck et al, 1997).

In part III, we consider the combination of bootstrap and Bayes inferences. In the case of independent identically distributed samples, the simple bootstrap yields confidence limits that are asymptotically correct to the first order but have less reliable confidence coverage in small samples. Bayesian credibility intervals based on the posterior distribution of the model parameters tend to perform better for small samples, but are more dependent on modeling assumptions than the bootstrap. A discrepancy statistic based on the difference of model and bootstrap estimates of variance is developed to combine bootstrap and Bayesian inferences. Our goal is to achieve a compromise that combines the advantages of those two methods, yielding intervals that combine robustness with good small-sample confidence coverage. We assess properties of our method by some simple simulation experiments which show some promise for the proposed method.

CHAPTER I

Introduction

In observational studies and clinical trials, missing data may arise for many reasons. For example, in a cross-sectional study relying on a survey, subjects may refuse to participate in the entire study or may not answer certain questions in the questionnaire. In a longitudinal study, participants may drop out from follow-up data collections. In a clinical trial testing the efficacy of a new drug, patients may not continue the trial due to severe side effects or other reasons. To yield efficient estimators and valid inferences, it is important to take account of the missing data in the analysis.

Many methods have been developed to deal with missing information. A simple approach is complete case (CC) analysis, which deletes units with any missing values, and therefore loses the information contained in the deleted cases. CC analysis is a default option in many statistical packages, however, it is inefficient and potentially biased, especially if the subjects included in the analysis are systematically different from those excluded in terms of one or more key variables. Another ad hoc approach is available case (AC) analysis, where restricts the analysis to the cases with variables of interest present. AC analysis uses all the available cases, its disadvantage is that the sample base changes from variable to variable according to the pattern of missing data.

With the loss of information contained in the deleted cases, both CC analysis and AC analysis yield less efficient estimators. To make full use of observed information, parametric approaches can be developed to deal with missing data, such as the maximum likelihood (ML) method, fully Bayesian (FB) method, and multiple imputation (MI). Unlike ad hoc approaches, parametric methods require an additional specification of a distribution for variables with missing values and/or the specification of the mechanisms that generate the missing values.

The ML method is based on the likelihood constructed from the observed incomplete data. This method has a long history: the earliest reference seems to be McKendrick (1926), where an algorithm similar to the Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin, 1977) is used to obtain estimates from a sample with missing values. The EM algorithm is a popular approach to maximizing the observed data likelihood. Each iteration of EM consists of an E step (expectation step) and an M step (maximization step). Each step has a direct statistical interpretation and is easy to construct conceptually and computationally. The EM algorithm is shown to converge reliably, in the sense that under general conditions, it converges to a local maximum or saddle point of the loglikelihood of the observed data, however, its convergence rate can be painfully slow when there is a large proportion of missing values. Another disadvantage of the EM algorithm is that the M step has no closed form in some problems. When the M step of the EM algorithm is iterative, several methods have been developed to modify the M step, such as the ECM (Expectation Conditional Maximization) algorithm (Meng and Rubin, 1993), the ECME (Expectation Conditional Maximization Either) algorithm (Liu and Rubin, 1994) and the AECM (Alternating Expectation Conditional Maximization) algorithm (Meng and van Dyk, 1997). A detailed discussion of the ML method and the EM algorithm can be found in Little and Rubin (2002). The

EM algorithm has broad applications. For example, it is useful to fit contingency tables with supplemental margins (Chen and Fienberg, 1974; Fuchs, 1982; Baker and Laird, 1985; Fay, 1986; Rubin, Stern and Vehovar, 1996; Little and Rubin, 2002, Section 15.7).

The FB method for missing data involves specifying priors for all parameters in the modeling as well as specifying the distribution for variables with missing values. The missing variables are then sampled from their posterior predictive distributions via Markov Chain Monte Carlo (MCMC) method or Gibbs' sampler. Compared to the case without missing variables, FB with missing variables needs to incorporate an extra layer in the Gibbs step. Thus, the Bayesian method can easily accommodate missing data without requiring new techniques for inference. The ML method is actually connected with the FB method, in that, ML can be viewed as a large sample Bayesian method and the Bayesian method using uniform priors on all parameters leads to ML estimates as modal Bayesian estimates.

MI originates from the Bayesian method. It involves creating multiple complete data sets by filling in values for the missing data. Then, each filled-in data set is analyzed as if it were a complete data set. The inferences for the filled-in data sets are then combined into one result by Rubin's combination rules (Rubin, 1978, 1987, 1996; Rubin and Schenker, 1986; Barnard and Rubin, 1999).

All these approaches need certain assumptions about why data are missing. The missing-data mechanism concerns whether and/or how the missingness depends on the values of variables in the data set. Let $Z = (Z_{ij})$ denote a rectangular $n \times p$ data set; the ith row is $Z_i = (Z_{i1}, \dots, Z_{ip})$, where Z_{ij} is the jth variable for subject i. Let $M = (M_{ij})$ be a missing data indicator matrix with the ith row $M_i = (M_{i1}, \dots, M_{ip})$, such that M_{ij} is 1 if Z_{ij} is missing and M_{ij} is 0 if Z_{ij} is present. We assume that

$(Z_i, M_i), i = 1, \ldots, n$ are independent throughout the paper. Rubin (1976a) treated M as a random matrix and described the missing-data mechanism by the conditional distribution of M given Z, say $f(M|Z, \psi)$, where ψ denotes unknown parameters. When missingness does not depend on the values of the data Z, missing or observed, that is, if

$$f(M|Z, \psi) = f(M|\psi) \text{ for all } Z, \psi,$$

the data are called missing completely at random (MCAR). If missingness depends only on the observed values Z_{obs}, and not on the missing values Z_{mis}. That is,

$$f(M|Z, \psi) = f(M|Z_{obs}, \psi) \text{ for all } Z_{mis}, \psi,$$

then the missing-data mechanism is called missing at random (MAR). If the distribution of M depends on the missing values in the data matrix Z, then the data are called not missing at random (NMAR). It is useful to distinguish the missing-data mechanism and the missing-data pattern, defined by M, which describes which values are observed and which values are missing in the data matrix. Many methods of handling missing data assume missingness is MCAR or MAR. If this is assumed, the missing-data mechanism can be ignored and we only need the observed data Z_{obs} to derive inferences. However, these inferences are subject to bias when the data are not MAR. In the dissertation, we consider the ML method for a NMAR mechanism we call almost missing at random (AMAR), which is close to MAR and realistic in some settings.

There are two ways to specify the joint distribution of Z and M. Selection models specify

$$p(Z, M|\theta, \psi) = p(Z|\theta)p(M|Z, \psi)$$

where $p(Z|\theta)$ and $p(M|Z, \psi)$ represent the models for complete data and the missing-data mechanism respectively, and θ is the parameter of interest. Pattern-mixture models specify

$$p(Z, M|\varphi, \pi) = p(Z|M, \varphi)p(M|\pi)$$

where Z is conditionally distributed on the missing-data pattern M, and (φ, π) are unknown parameters. Noting that the resulting marginal distribution of Z is a mixture of distributions, Glynn, Laird and Rubin (1986, 1993) used the term "mixture" for this kind of model, while Little (1993, 1994) used term "pattern-mixture" to reflect these models, where "pattern" is added to make the nature of the mixing more explicit. When data are MCAR, these two specifications are equivalent if $\theta = \varphi$ and $\psi = \pi$. When data are not MCAR, they can yield different models providing additional assumptions are added. ML for selection models requires numerical methods such as EM algorithm, whereas additional conditions are often needed to identify pattern-mixture models.

For missing data with a monotone pattern, where variables can be arranged so that $Z_{j+1}, \ldots, Z_P$ are missing for cases with Z_j missing, for all $j = 1, \ldots, P-1$, the definition of MAR is intuitive. For example, in longitudinal studies collecting information on a set of subjects repeatedly over time, a common reason for missing data is attrition, where subjects drop out prior to the end of the study and do not return. The missing-data mechanism is MAR provided the missingness of Z_j depends only on the previous recorded history, that is on observed values of $(Z_1, ..., Z_{j-1})$ (and fully observed baseline covariates, if they exist). Methods for handling monotone missing data can be easier than methods for general patterns, as shown in Little and Rubin (2002, Chapter 7) where inferences are made from factored likelihood methods. When the data do not have a monotone pattern, many existing analyses assume MAR, thus they do not need to additionally model the missing-data mechanism to derive the inferences (Ibrahim, 1990; Ibrahim, Chen and Lipsitz, 1999; Lipsitz and Ibrahim, 1996; Lipsitz, Ibrahim, and Fitzmaurice, 1999; Stubbendick and Ibrahim, 2003). However, it is less intuitive to define MAR for general pattern missing data. For example, in longitudinal data with two variables Z_1 and Z_2, where Z_1 is recorded followed by Z_2, and both of two variables have missing values, MAR assumes that missingness of Z_2 given that Z_1 is observed depends only on Z_1 and missingness of Z_1 given that Z_2 is observed depends only on the value of Z_2. Since the missingness of Z_1 depends on Z_2 measured at a later time, it is not intuitive from a causal perspective (Little and Rubin, 2002, Example 1.13). In the dissertation, we consider an interesting missing-data mechanism that is NMAR but is similar in some respects to MAR mechanisms for two variables with a general missing-data pattern. We assume Z_1 is MCAR, and the missingness of following Z_2 depends on the value of Z_1 and on whether Z_1 is missing or not. Since the missingness of Z_2 can depend on the value of Z_1 even when it is not observed, this missing-data mechanism is NMAR. If Z_1 were fully observed, it would be MAR. For this reason, we refer to this missing-data mechanism as almost MAR, or AMAR.

There appears to be very little existing literature on missing-data mechanisms of the type considered here; most of the work on NMAR mechanisms concerns the situation where missingness depends directly on outcomes of interest, or on latent variables such as the slope of a repeatedly measured variable (e.g. Little and Rubin 2002, chapter 15). Perhaps the closest work to that presented here is latent ignorable missing-data mechanisms proposed by Frangakis and Rubin (1999) to model missing data in a randomized clinical trial with noncompliance to the treatment assignments.

Non-compliance is a common issue in randomized clinical trails involving human subjects. It is often associated with the effects of treatments, and may vary according to participant characteristics. For example, in psychiatric trials, subjects' mental health conditions may affect their ability or willingness to comply with study protocols. A standard intention-to-treat (ITT) analysis compares the difference in outcome distributions based on treatment assignments (T). By ignoring non-compliance information, it only provide a valid measurement of the effect of the treatment assignment, not treatment efficacy which is the effect of the treatment itself. To estimate treatment efficacy, information of non-compliance has to be incorporated into the analysis. Here we consider principal compliance instead of observed compliance, which concerns only whether a participant complied with the treatment assignment. Principal compliance (C) is a special case of principal stratification (Frangakis and Rubin 2002), where individuals are stratified according to the values of the post-treatment variable (such as compliance) under both treatments, rather than simply under the treatment actually assigned. Principal compliance stratifies the population into three groups. Compliers who take their assigned treatment, never-takers who take the control treatment no matter which treatment they are assigned, always-takers who take the active treatment whether they are assigned the active or control treatment. We assume there are no defiers who take the opposite to the treatment assigned. In the two-arm (active treatment vs. control treatment) randomized clinical trials we consider, participants in the active treatment group may switch to take the control treatment, therefore, they are observed to be either principal compliers or never-takers. On the other hand, participants in the control group don't have access to the active treatment, and whether they are principal compliers or never-takers is not observed/unknown. Besides missing principal compliance for those in the control group, analyzing randomized clinical trials may be further complicated with consequent missing outcomes due to loss to follow-up or non-response. Researchers have only recently started to develop methods for handling both non-compliance and subsequent nonresponses in the same study (Frangakis and Rubin, 1999; Levy, O'Malley and Normand, 2004; O'Malley and Normand, 2005; Peng, Little and Raghunathan, 2004; Zhou and Li, 2006). Under assumptions of latent ignorability (outcomes are missing at random conditional on latent compliance status and treatment assignments) and compound exclusion restrictions (the missingness and potential values of outcomes are independent of treatment assignments for never-takers), Frangakis and Rubin (1999) proposed a method-of-moments (MOM) estimator. Under the same assumptions, Zhou and Li (2006) derived ML estimates when the outcome is binary and O'Malley and Normand (2005) obtained ML estimators for normal distributed outcomes using an EM algorithm.

In chapter II, we examine in some detail a special case of AMAR for bivariate categorical data assumed to have a multinomial distribution. EM can always be implemented to seek the ML estimates, however, when the number of levels of Z_1 is equal to or greater than that of Z_2 and the closed form 'estimates' of certain nuisance parameters lie inside their admissible range $[0, 1]$, non-iterative ML estimates (obtained from the pattern-mixture model) exist, with some data being excluded for estimating the parameters of interest. We also introduce a restricted version of the AMAR model where the missingness of Z_2 depends on Z_1 but not on the missingness of Z_1. We present some numerical examples to illustrate when explicit ML estimates exist for the parameters in the AMAR models and apply the AMAR models to data from the Muscatine coronary risk factor study (Woolson and Clark, 1984).

In chapter III, we consider one extension of AMAR. Besides two variables with missing values, there is an additional fully observed covariate. Specifically, we consider randomized clinical trials with non-compliance to the treatment assignments and subsequent non-response (Y). We build a connection between latent ignorability and AMAR. To identify the model, we further specify two models for principal compliance, ER (exclusion restriction which indicates there is no effect of T on the distribution of Y for never-takers) or NCEC (none compliance effect in controls which implies the distribution of Y is same for compliers and never-takers in the control group), and two models for missing outcome, ER (there is no effect of T

on missingness of Y for never-takers) or NCEC (there is no effect of C on missingness of Y for participants in the control group). We consider all four combinations for a clinical trial with a categorical outcome. Both complier average causal effect (CACE) or ITT estimands can be viewed as outputs from these models. By applying the pattern-mixture model with covariates (Little and Wang, 1996), we show non-iterative ML estimates sometimes exist in each combination. We find the models of principal compliance determine which analysis is used to estimate treatment efficacy, such as per-protocol analysis or IV estimation with the treatment assignment indicator as the instrumental variable, whereas the models of missing outcome decide whether non-iterative ML estimates exist or not. We apply our methods to analyze the data from a double-blinded randomized clinical trials with clozapine vs. haloperidol (Rosenheck et al, 1997).

In chapter IV, we consider a different topic in adaptive confidence intervals. By eliminating the routine but tedious theoretical calculations usually associated with precision assessment, Bootstrap methods (e.g. Efron, 1979, 1981, 1982) provide tools that can be used to set confidence intervals in complex problems. However, they yield confidence limits that are asymptotically correct to the first order, therefore perform poorly in some small sample problems, such as setting a confidence interval for the variance (Schenker, 1985). Bayesian credibility intervals based on the posterior distribution of the model parameters tend to perform better for small samples, but are more dependent on modeling assumptions than the bootstrap. In this chapter, based on the difference of model and bootstrap estimates of variance, we introduce a discrepancy statistic and construct a function of its posterior predictive p-value (Guttman, 1967; Rubin, 1981, 1984) to combine bootstrap and Bayesian inferences. The goal is to achieve a compromise that combines the advantages of those two methods, yielding intervals with robustness and good small-sample confidence coverage. We assess properties of our method by some simple simulation experiments. We conclude the dissertation with a short discussion and future work in chapter V.

CHAPTER II

Likelihood Method for Data with Almost MAR Mechanisms

Abstract EM is a simple and intuitive algorithm for maximum likelihood estimation for contingency tables with missing data. The missing data mechanism can be ignored when the data are missing at random, but this assumption is not always intuitive for general patterns of missing data. We consider maximum likelihood (ML) estimation for a nonignorable mechanism we call almost missing at random (AMAR), which is close to missing at random and realistic in some settings. We examine in some detail the case of two multinomially distributed categorical variables X and Y, for which X is missing completely at random and Y is MAR given the value and missingness of X. In this case, ML can be fitted using EM when ML estimates are at the boundary of the parameter space, but otherwise (rather surprisingly) non-iterative ML estimates exist, with some data being excluded for estimating the parameters of interest. Extensions of this type of mechanism are also discussed. We apply the AMAR models to data from the Muscatine coronary risk factor study.

keywords: missing data, EM algorithm, categorical data, almost MAR mechanism.

2.1 Introduction

Missing values arise in empirical studies for many reasons, including the unavailability of the measurements, survey nonresponse, respondents refusing to answer certain items on a questionnaire, and attrition in longitudinal studies. Complete case (CC) analysis, which omits information in the cases with missing values, is inefficient and potentially biased, especially if the subjects included in the analysis are systematically different from those excluded in terms of one or more key variables. Approaches that incorporate information in the incomplete cases include nonresponse weighting (Little and Rubin 2002, chapter 3); multiple imputation (MI), where missing values are replaced by several plausible values (Rubin 1987; Little and Rubin 2002, chapter 5); weighted estimating equation (WEE) methods (Lipsitz, Ibrahim and Zhao, 1999); and methods based on the likelihood for a model for the data, such as maximum likelihood (ML) or fully Bayes modeling. We focus here on the ML approach.

The performance of alternative missing-data methods depends on the missing-data mechanism, which concerns why values are missing, and in particular, whether the missingness depends on the values of variables in the data set. Rubin (1976a) formalized the concept of missing-data mechanisms by treating the missing-data indicators as random variables and assigning them a distribution. Let $Z = (Z_{ij})$ denote a rectangular $n \times p$ data set; the ith row is $Z_i = (Z_{i1}, \dots, Z_{ip})$, where Z_{ij} is the jth observation for subject i. Let $M = (M_{ij})$ be a missing data indicator matrix with the ith row $M_i = (M_{i1}, \dots, M_{ip})$, such that M_{ij} is 1 if Z_{ij} is missing and M_{ij} is 0 if Z_{ij} is present. We assume that $(Z_i, M_i), i = 1, \dots, n$ are independent throughout the paper. The missing-data mechanism is then characterized by the conditional distribution of M given Z, say $f(M|Z, \psi)$, where ψ denotes unknown parameters. When missingness does not depend on the values of the data Z, missing or observed, that is, if

$$f(M|Z,\psi) = f(M|\psi) \text{ for all } Z,\, \psi,$$

the data are called missing completely at random (MCAR). With the exception of planned missing-data designs, MCAR is a strong assumption, and missingness often depends on the observed (or unobserved) data. Let Z_{obs} denote the observed component of Z and Z_{mis} the missing component. A less restrictive assumption is that missingness depends only on the observed values Z_{obs}, and not on the missing values Z_{mis}. That is,

$$f(M|Z,\psi) = f(M|Z_{obs},\psi) \text{ for all } Z_{mis},\, \psi..$$

The missing-data mechanism is then called missing at random (MAR). The mechanism is called not missing at random (NMAR) if the distribution of M depends on the missing values in the data matrix Z.

In general, the actual observed data consist of the values of the variables (Z_{obs}, M) and the distribution of the observed data is obtained by integrating Z_{mis} out of the joint density of $Z = (Z_{obs}, Z_{mis})$ and M. That is,

$$f(Z_{obs}, M|\theta,\psi) = \int f(Z_{obs}, Z_{mis}|\theta) f(M|Z_{obs}, Z_{mis},\psi) dZ_{mis}. \tag{2.1}$$

where θ is the vector of parameters in the distribution of Z to be estimated. The full likelihood of θ and ψ is any function of θ and ψ proportional to (2.1):

$$L_{full}(\theta,\psi|Z_{obs}, M) \propto f(Z_{obs}, M|\theta,\psi).$$

If the missing mechanism is ignorable, that is, if the mechanism is MAR and θ and ψ are distinct, in the sense that $(\theta,\psi) \in \Theta \times \Psi$ where Θ and Ψ are parameter spaces, then likelihood based inferences for θ from $L_{full}(\theta,\psi|Z_{obs}, M)$ will be the same as likelihood based inferences for θ from $L_{ign}(\theta|Z_{obs})$, the likelihood of θ based on the observed data Z_{obs} (Rubin, 1976a). Many methods of handling missing data assume missingness is MCAR or MAR. If this is assumed, the missing-data mechanism can be ignored and we only need the observed data Z_{obs} to derive the likelihood-based inferences for θ. However, these inferences are subject to bias when the data are not MAR.

The focus of this chapter is on ML methods for categorical Z where the complete cases form a p-way contingency table, and the incomplete cases form supplemental margins (see for example Little and Rubin 2002, Chapter 13). The EM algorithm, the topic of this special issue, is particularly appealing for incomplete categorical data, since the natural distributions for modeling count data, the Poisson and multinomial distributions, yield complete data loglikelihoods that are in the exponential family and are linear in the cell counts. Consequently, the E step of EM consists of replacing the complete-data cell counts by conditional expectations given the observed data, in effect distributing the supplemental margins into the full table according to current estimates of the cell probabilities. The M step of EM is the same as complete-data ML estimation based on the data filled in by the E step. This approach to estimation for count data with some grouped counts was first established as ML by Hartley (1958). The application to a (2x2) table with supplemental margins was considered by Chen and Fienberg (1974), and extended to the general class of loglinear models by Fuchs (1982).

When the M step of EM is iterative, standard EM involves a double iteration, with the M step being achieved by the Deming Stephan algorithm, otherwise known as iterative proportional fitting (e.g. Bishop, Fienberg and Holland, 1975). If the M step is restricted to just one iteration of Deming-Stephan, the likelihood function is increased, and hence the result is an example of the ECM algorithm (Meng and Rubin, 1993), a form of generalized EM algorithm that shares similar theoretical properties to EM with a single iterative loop. EM is also useful for fitting nonignorable models for contingency tables (Baker and Laird, 1985; Fay, 1986: Rubin, Stern and Vehovar, 1996; Little and Rubin, 2002, Section 15.7). In this article we present ML results for some interesting missing data mechanisms that are nonignorable but are similar in some respects to MAR mechanisms. ML for these models is sometimes noniterative, but can be fitted using EM when ML estimates are at the boundary of the parameter space.

The definition of MAR is intuitive for monotone patterns, where variables can be arranged so that Z_{j-1} is observed whenever Z_j is observed, for all $j = 1, \ldots, p$. An important example is longitudinal data subject to attrition, where the mechanism is MAR provided the missingness of Z_j depends only on the previous recorded history, that is on observed values of $(Z_1, ..., Z_{j-1})$ (and fully observed baseline covariates, if they exist). The point of departure for our work is the observation that when the data do not have a monotone pattern, the MAR definition is less intuitive. For example, consider longitudinal data on two variables Z_1 and Z_2, where Z_1 is recorded and then Z_2, and both Z_1 and Z_2 have missing values; MAR corresponds to the assumption that missingness of Z_2 given that Z_1 is observed depends only on Z_1 and missingness of Z_1 given that Z_2 is observed depends only on the value of Z_2. The latter is not intuitive from a causal perspective, since it implies that missingness of Z_1 depends on a variable measured at a later time (Little and Rubin, 2002, Example 1.13).

Other mechanisms may correspond more closely to our intuitive notion of random missingness. In particular, we consider here the situation, again with bivariate data (Z_1, Z_2) where Z_1 is MCAR, and missingness of Z_2 depends on the value of Z_1 and on whether Z_1 is missing. Although this mechanism seems MAR-like, it does not meet the formal definition of MAR, since missingness of Z_2 can depend on the value of Z_1 even when it is not observed. For want of a better label, we call this mechanism "almost MAR"(AMAR), since it would be MAR if Z_1 were fully observed, and the missing values of Z_1 are themselves MCAR.

In section 2.2, we consider ML estimation for this AMAR mechanism, for the special case of bivariate categorical data assumed to have a multinomial distribution. The results are surprising. In particular, we show that in many situations, explicit ML estimates are available that exclude the data with Z_1 missing and estimate the parameters of the joint distribution of (Z_1, Z_2) from the resulting monotone pattern. However, when the closed form 'estimates' of certain nuisance parameters lie outside their admissible range $[0, 1]$, the data with Z_1 missing enter into the estimation! In section 2.3, a restricted version of the AMAR model is introduced where missingness of Z_2 depends on Z_1 but not on whether Z_1 is missing. Some numerical examples are presented in section 2.4 to illustrate when explicit ML estimates exist for the parameters in the AMAR models. A real data example is given in section 2.5, and some concluding remarks on extensions of this AMAR model are made in section 2.6.

2.2 Unrestricted AMAR model

We consider data where X and Y are categorical variables respectively with J and K categories. Both X and Y may be missing, so there are four missing-data patterns. Let $r = 0, 1, 2, 3$ index the missing-data patterns and let P_r denote the set of sample cases with pattern type r, $r = 0, \ldots, 3$ (see Table 2.1). Let n_r denote the number of cases in the sample with pattern r and $n = \sum_r n_r$ denote the total sample size.

For categorical X and Y with J and K levels, data in P_0 can be arranged as a $J \times K$ contingency table, and the data in P_1 and P_2 form supplemental $J \times 1$ and $1 \times K$ margins. Let $n_{(0),jk}$ be the count of complete cases with $X = j, Y = k$, $n_{(1),j+}$ be the

Table 2.1: Missing-Data Pattern for Two Variables

Pattern		
P_0	■	■
P_1	■	?
P_2	?	■
P_3	?	?

count of cases with $X = j$ and Y missing, $n_{(2),+k}$ be the count of cases with $Y = k$ and X missing, and $n_{(3),++}$ be the count of cases with both X and Y missing. The data are displayed in Table 2.2. Note that $n_0 = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{(0),jk}$, $n_1 = \sum_{j=1}^{J} n_{(1),j+}$, $n_2 = \sum_{k=1}^{K} n_{(2),+k}$, and $n_3 = n_{(3),++}$.

Table 2.2: Notation for a J×K Table with Supplemental Margins for Both Variables

		Y					
		1	2	...	...	K	missing
X	1	$n_{(0),11}$	$n_{(0),12}$	...	...	$n_{(0),1K}$	$n_{(1),1+}$
	2	$n_{(0),21}$	$n_{(0),22}$	...	...	$n_{(0),2K}$	$n_{(1),2+}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	J	$n_{(0),J1}$	$n_{(0),J2}$	...	...	$n_{(0),JK}$	$n_{(1),J+}$
	missing	$n_{(2),+1}$	$n_{(2),+2}$	...	...	$n_{(2),+K}$	$n_{(3),++}$

The parameters of interest are $\theta = \{\theta_{jk}\}$, where $\theta_{jk} = P(X = j, Y = k)$ with $\sum_{j=1}^{J} \sum_{k=1}^{K} \theta_{jk} = 1$. The MAR assumption for these data implies that

$$P(M^X = M^Y = 1|X = j, Y = k) = \upsilon,$$

$$P(M^Y = 1, M^X = 0|X = j, Y = k) = \upsilon_j^{(0)},$$

$$P(M^X = 1, M^Y = 0|X = j, Y = k) = \upsilon_k^{(1)},$$

$$P(M^X = M^Y = 0|X = j, Y = k) = 1 - \upsilon - \upsilon_j^{(0)} - \upsilon_k^{(1)}.$$

where $1 \leq j \leq J$, $1 \leq k \leq K$ (See Little and Rubin 2002, Example 1.19). In this case, $\upsilon = \{\upsilon, \upsilon_j^{(0)}, \upsilon_k^{(1)}\}$ represent nuisance parameters for the missing-data mechanism. Under MAR, the likelihood factors into distinct components of θ and υ; ML estimation of θ under MAR involves all the observed data and requires an iterative algorithm such as EM (Little and Rubin 2002, Chapter 13).

We consider as an alternative to MAR the following AMAR model, which incorporates the assumption that X is MCAR and missingness of Y depends on X and M^X:

$$\begin{aligned} &P(M^X = 1|X = j, Y = k) = \phi, \\ &P(M^Y = 1|M^X = 0, X = j, Y = k) = \phi_j^{(0)}, \\ &P(M^Y = 1|M^X = 1, X = j, Y = k) = \phi_j^{(1)}. \end{aligned} \tag{2.2}$$

where $1 \le j \le J$, $1 \le k \le K$. Here $\psi = \{\phi, \phi_j^{(0)}, \phi_j^{(1)}\}$ are nuisance parameters corresponding to the missing-data mechanism. The number of the parameters in this model is $JK + 2J$, whereas the degrees of freedom of the data are $JK + J + K$, which comprises JK for the complete cases, plus J for the supplemental margin on X, plus K for the supplemental margin on Y, plus 1 for the number of cases with X and Y both missing, minus 1 for the total which is considered fixed at n. When $J = K$, the model has the same number of parameters as degrees of freedom in the data; otherwise, the model has more parameters for $J > K$ or fewer for $J < K$.

Note that if $\phi_j^{(1)} = \phi^{(1)}$ does not depend on j, this reduces to a restricted MAR model in which X is MCAR and missingness of Y may depend on the observed values of X and M^X. A likelihood ratio test could be used to test this restricted MAR assumption against the more general AMAR model. Another submodel of interest is discussed in Section 3.

2.2.1 EM algorithm

The likelihood for the above model has the form:

$$L(\theta, \psi|X_{obs}, Y_{obs}, M) = \sum_{X_{mis}} \sum_{Y_{mis}} \left\{ \prod_{i=1}^{n} p(X_i, Y_i|\theta) p(M_i^X|X_i, Y_i, \phi) \right. \\ \left. p(M_i^Y|M_i^X, X_i, Y_i, \phi_j^{(0)}, \phi_j^{(1)}) \right\}$$

$$\begin{aligned} = \sum_{X_{mis}} \sum_{Y_{mis}} \Bigg\{ & \prod_{i=1}^{n} \prod_{j,\,k=1}^{J,\,K} \theta_{jk}^{I(X_i=j,Y_i=k)} \phi^{I(M_i^X=1)} (1-\phi)^{I(M_i^X=0)} \\ & \prod_{j=1}^{J} {\phi_j^{(0)}}^{I(M_i^X=0,X_i=j,M_i^Y=1)} (1-\phi_j^{(0)})^{I(M_i^X=0,X_i=j,M_i^Y=0)} \\ & \prod_{j=1}^{J} {\phi_j^{(1)}}^{I(M_i^X=1,X_i=j,M_i^Y=1)} (1-\phi_j^{(1)})^{I(M_i^X=1,X_i=j,M_i^Y=0)} \Bigg\}. \end{aligned}$$

where $I(.)$ is the indicator function and M_i^X, M_i^Y are the missing indicators for variable X and Y in case i respectively.

As for the general MAR mechanism, one approach to ML estimation is to apply the EM algorithm (Dempster, Laird and Rubin, 1977). To define the E step of EM, let $(\theta_{jk}^{(t)}, {\phi_j^{(1)}}^{(t)})$ denote the parameter estimates at iteration t, and $n_{(r),jk}^{(t)}$ be the estimate of cell frequency for $X=j, Y=k$ in pattern P_r. The E step distributes the partially classified observations into the table according to the corresponding probabilities:

$$\begin{aligned} n_{(1),jk}^{(t)} = &\quad n_{(1),j+} \cdot \frac{\theta_{jk}^{(t)}}{\theta_{j+}^{(t)}} , \\ n_{(2),jk}^{(t)} = &\quad n_{(2),+k} \cdot \frac{(1-{\phi_j^{(1)}}^{(t)})\theta_{jk}^{(t)}}{\sum_{j=1}^{J}(1-{\phi_j^{(1)}}^{(t)})\theta_{jk}^{(t)}} , \\ n_{(3),jk}^{(t)} = &\quad n_{(3),++} \cdot \frac{{\phi_j^{(1)}}^{(t)}\theta_{jk}^{(t)}}{\sum_{j=1}^{J}{\phi_j^{(1)}}^{(t)}\theta_{j+}^{(t)}} . \end{aligned}$$

The M step calculates new parameters as:

$$\begin{aligned} \theta_{jk}^{(t+1)} = &\quad \frac{n_{(0),jk} + n_{(1),jk}^{(t)} + n_{(2),jk}^{(t)} + n_{(3),jk}^{(t)}}{n} , \\ \phi^{(t+1)} = &\quad \frac{\sum_{i=1}^{n} I(M_i^X = 1)}{n} = \frac{n_2 + n_3}{n} , \\ {\phi_j^{(0)}}^{(t+1)} = &\quad \frac{\sum_{i=1}^{n} I(M_i^X = 0, M_i^Y = 1, X_i = j)}{\sum_{i=1}^{n} I(M_i^X = 0, X_i = j)} = \frac{n_{(1),j+}}{n_{(1),j+} + n_{(0),j+}} , \\ {\phi_j^{(1)}}^{(t+1)} = &\quad \frac{\sum_k n_{(3),jk}^{(t)}}{\sum_k n_{(2),jk}^{(t)} + \sum_k n_{(3),jk}^{(t)}} . \end{aligned}$$

The E-step and M-step alternate until the parameter estimates converge.

Note that ϕ and $\{\phi_j^{(0)}\}$ are estimable directly and are unchanged throughout the EM algorithm. Complete-case estimates or estimates arising from the monotone pattern P_0 and P_1 can be chosen as the starting values of $\{\theta_{jk}\}$, and the estimates of $\{\phi_j^{(0)}\}$ or any constant in $(0, 1)$ can be taken as initial values of $\{\phi_j^{(1)}\}$. When $J > K$, the model has more parameters than degrees of the freedom in the data. Multiple maxima may exist in this case. Depending on starting values, the algorithm can converge to different estimates. This case will be discussed further below.

2.2.2 Non-iterative ML estimates

When $J \geq K$, non-iterative estimates of the parameters can sometimes be obtained using the factored likelihood method (Little & Rubin 2002, chapter 7). We transform the parameters $(\theta_{jk}, \phi, \phi_j^{(0)}, \phi_j^{(1)})$ to:

$$\begin{aligned}
&\alpha_{(0),jk} = P(X = j, Y = k|M^X = M^Y = 0),\\
&\alpha_{(1),j+} = P(X = j|M^X = 0, M^Y = 1),\\
&\alpha_{(2),+k} = P(Y = k|M^X = 1, M^Y = 0),\\
&\pi_0 = P(M^X = 0, M^Y = 0),\ \pi_1 = P(M^X = 0, M^Y = 1),\\
&\pi_2 = P(M^X = 1, M^Y = 0),\ \pi_3 = P(M^X = 1, M^Y = 1).
\end{aligned} \tag{2.3}$$

where $1 \leq j \leq J$, $1 \leq k \leq K$ and the following constraints apply:

$$\sum_{j=1}^{J}\sum_{k=1}^{K}\alpha_{(0),jk} = 1,\ \sum_{j=1}^{J}\alpha_{(1),j+} = 1,\ \sum_{k=1}^{K}\alpha_{(2),+k} = 1,\ \sum_{r=0}^{3}\pi_r = 1.$$

These parameters correspond to a pattern-mixture factorization (Little, 1993):

$$f(X_{obs}, Y_{obs}, M|\alpha, \pi) = f(X_{obs}, Y_{obs}|M, \alpha)f(M|\pi).$$

where $\alpha = (\alpha_{(0),jk}, \alpha_{(1),j+}, \alpha_{(2),+k})$, and $\pi = (\pi_r)$.

The components of $(\theta, \psi) = (\theta_{jk}, \phi, \phi_j^{(0)}, \phi_j^{(1)})$ can be expressed in terms of the new parametrization (2.3) as follows:

$$\begin{aligned}
\theta_{jk} &= \frac{\alpha_{(0),jk}}{\alpha_{(0),j+}} \cdot \frac{\pi_0 \alpha_{(0),j+} + \pi_1 \alpha_{(1),j+}}{\pi_0 + \pi_1} \ , \\
\phi &= 1 - \pi_0 - \pi_1 \ , \\
\phi_j^{(0)} &= \frac{\pi_1 \alpha_{(1),j+}}{\pi_0 \alpha_{(0),j+} + \pi_1 \alpha_{(1),j+}} \ ,
\end{aligned} \tag{2.4}$$

and $\{\phi_j^{(1)},\ j = 1, ..., J\}$ is a solution to the K simultaneous equations

$$\sum_{j=1}^{J} (1 - \phi_j^{(1)}) \theta_{jk} = P(M^Y = 0, Y = k | M^X = 1) = \frac{\pi_2}{1 - \pi_0 - \pi_1} \alpha_{(2),+k} \ .$$

where $\alpha_{(0),j+} = \sum_{k=1}^{K} \alpha_{(0),jk}$.

Under pattern-mixture factorization, the likelihood can be written as

$$\begin{aligned}
L(\varphi, \pi | X_{obs}, Y_{obs}, M) &= \prod_{i=1}^{n} p(M_i^X, M_i^Y) \prod_{i \in p_0} p(X_i, Y_i | M_i^X = 0, M_i^Y = 0) \\
&\times \prod_{i \in p_1} p(X_i | M_i^X = 0, M_i^Y = 1) \prod_{i \in p_2} p(Y_i | M_i^X = 1, M_i^Y = 0) \\
&= \prod_{r=0}^{3} \pi_r^{n_r} \prod_{j,\, k=1}^{J,\, K} \alpha_{(0),jk}^{n_{(0),jk}} \prod_{j=1}^{J} \alpha_{(1),j+}^{n_{(1),j+}} \prod_{k=1}^{K} \alpha_{(2),+k}^{n_{(2),+k}} \ .
\end{aligned}$$

Maximizing the four terms in this likelihood yields

$$\hat{\alpha}_{(0),jk} = \frac{n_{(0),jk}}{n_0} \ , \qquad \hat{\alpha}_{(1),j+} = \frac{n_{(1),j+}}{n_1} \ , \qquad \hat{\alpha}_{(2),+k} = \frac{n_{(2),+k}}{n_2} \ , \qquad \hat{\pi}_r = \frac{n_r}{n} \ . \tag{2.5}$$

where $1 \leq j \leq J$, $1 \leq k \leq K$ and $0 \leq r \leq 3$. Estimates of θ_{jk}, ϕ and $\phi_j^{(0)}$ can then be obtained by substituting the above estimates of $(\alpha, \pi) = (\alpha_{(0),jk}, \alpha_{(1),j+}, \alpha_{(2),+k}, \pi_r)$ into the equations (2.4). This yields:

$$\hat{\theta}_{jk} = \left(\frac{n_{(0),jk}}{n_{(0),j+}} \right) \left(\frac{n_{(0),j+} + n_{(1),j+}}{n_0 + n_1} \right) \ , \tag{2.6}$$

$$\hat{\phi} = 1 - \hat{\pi}_0 - \hat{\pi}_1 \ , \tag{2.7}$$

$$\hat{\phi}_j^{(0)} = \frac{\hat{\pi}_1 \hat{\alpha}_{(1),j+}}{\hat{\pi}_0 \hat{\alpha}_{(0),j+} + \hat{\pi}_1 \hat{\alpha}_{(1),j+}} \ , \tag{2.8}$$

Estimates of $\{\phi_j^{(1)}, j = 1, ..., J\}$ can be obtained as solutions of the following K simultaneous equations, provided they are in the parameter space:

$$\sum_{j=1}^{J}(1 - \hat{\phi}_j^{(1)})\hat{\theta}_{jk} = \frac{\hat{\pi}_2}{1 - \hat{\pi}_0 - \hat{\pi}_1}\hat{\alpha}_{(2),+k} \ . \tag{2.9}$$

This approach yields ML estimates, providing the estimates lie within the parameter space, that is the probabilities lie between zero and one. The expressions for $\hat{\theta}_{jk}, \hat{\phi}$ and $\hat{\phi}_j^{(0)}$ always yield estimates in $[0, 1]$. The equations in (2.9), however, may or may not yield solutions for $\{\phi_j^{(1)}\}$ that lie in $[0, 1]$. If they do, then estimates from this approach are ML estimates. If not, this approach fails to yield ML estimates of the parameters of interest. The EM algorithm can still be used. The solution set for (2.9) depends on whether $J = K$ or $J > K$. When $J = K$ there are J equations for J unknowns. Provided the $J \times J$ matrix, $\hat{\Theta} = (\hat{\theta}_{jk})$, is non-singular, these equations yield a unique solution that may or may not lie in the parameter space. When $J > K$ and $\hat{\Theta}$ has rank K', the solution set is a linear subspace of dimension $J - K'$. If the solution space intersects the parameter space $[0, 1]^J$, then this approach yields the whole class of ML estimates. For example, consider the case where $J = 3$, $K = 2$ and $\hat{\Theta}$ is of full rank K, the solution set to (2.9) is a straight line. When it intersects with the unit cube which is the parameter space, this approach yields unique ML estimates of θ_{jk}, ϕ and $\phi_j^{(0)}$, although there are multiple ML estimates for $\{\phi_j^{(1)}\}$. However, when the straight line does not intersect with the unit cube, the EM algorithm can be implemented to find ML estimates, that may or may not be unique.

The closed-form estimates (2.6) of θ are simply the product of the estimated conditional probabilities of $Y = k$ given $X = j$ from the complete cases and the marginal probabilities of $X = j$ from the cases with X observed. Remarkably, they do not involve the data for Y from the pattern with Y observed and X missing, which one would expect to provide information for the marginal distribution of Y. However, under the model, the distribution of Y for cases in this pattern is different from the marginal distribution of Y overall. Dropping these cases and using just patterns P_0 and P_1 with a MAR analysis yields estimates that are ML, and hence asymptotically consistent and efficient, under the assumed model provided the solutions of Eq. (2.9) are interior to the parameter space.

2.3 A restricted AMAR Model

In the unrestricted AMAR model (2.2), the missingness of Y is allowed to depend not only on the value of X but also on whether X is missing or not. If, given the value of X, the probability of Y being missing is assumed the same for the cases with X observed and missing, we then have the restricted AMAR model:

$$\begin{aligned} &P(M^X = 1|X = j, Y = k) = \phi \ , \\ &P(M^Y = 1|M^X = l, X = j, Y = k) = \phi_j \ . \end{aligned} \tag{2.10}$$

where $l = 1, 2$ and $1 \leq j \leq J$, $1 \leq k \leq K$. The number of the parameters in this model is $JK + J$ which is always less than the degree of freedom $JK + J + K$ in the data. The explicit estimates in (2.6) are no longer ML estimates of $\{\theta_{jk}\}$, and EM is needed to obtain ML estimates of the parameters. In the E step, the partially classified observations are effectively distributed into the table according to the corresponding estimated probabilities:

$$\begin{aligned} n_{(1),jk}^{(t)} &= \quad n_{(1),j+} \cdot \frac{\theta_{jk}^{(t)}}{\theta_{j+}^{(t)}} \ , \\ n_{(2),jk}^{(t)} &= \quad n_{(2),+k} \cdot \frac{(1-{\phi_j}^{(t)})\theta_{jk}^{(t)}}{\sum_{j=1}^{J}(1-{\phi_j}^{(t)})\theta_{jk}^{(t)}} \ , \\ n_{(3),jk}^{(t)} &= \quad n_{(3),++} \cdot \frac{{\phi_j}^{(t)}\theta_{jk}^{(t)}}{\sum_{j=1}^{J}{\phi_j}^{(t)}\theta_{j+}^{(t)}} \ . \end{aligned}$$

In the M step, new estimates are calculated as:

$$\begin{aligned}
\theta_{jk}^{(t+1)} &= \frac{n_{(0),jk} + n_{(1),jk}^{(t)} + n_{(2),jk}^{(t)} + n_{(3),jk}^{(t)}}{n} , \\
\phi^{(t+1)} &= \frac{n_2 + n_3}{n} , \\
\phi_j{}^{(t+1)} &= \frac{\sum_k n_{(1),jk}^{(t)} + \sum_k n_{(3),jk}^{(t)}}{n_{(0),j+} + \sum_k n_{(1),jk}^{(t)} + \sum_k n_{(2),jk}^{(t)} + \sum_k n_{(3),jk}^{(t)}} .
\end{aligned}$$

The E-step and M-step alternate until the parameter estimates converge. Since ϕ is estimable directly and is unchanged throughout the EM algorithm, starting values are only needed for $\{\theta_{jk}\}$ and $\{\phi_j\}$. Complete-case estimates or pooled estimates arising from the monotone pattern P_0 and P_1 can be selected as the starting values of $\{\theta_{jk}\}$, and the estimates of $\{\phi_j^{(0)}\}$ in (2.8) or any constant in $(0, 1)$ can be taken as initial values of $\{\phi_j\}$.

With $\phi_j^{(0)} = \phi_j^{(1)}$, the restricted AMAR model (2.10) is a submodel of the unrestricted AMAR model (2.2). A likelihood ratio test can be applied to test the restricted AMAR assumption against the more general unrestricted AMAR model.

2.4 Numerical examples

2.4.1 Examples with $J = K = 2$

In Table 2.3, 3A gives data for a 2 × 2 table with supplemental margins. Estimates of $\{\phi_j^{(1)}\}$ from (2.9) lie in the parameter space, so there are closed form ML estimates under the unrestricted AMAR model (Table 2.4). For data in table 3B, 'estimates' of $\{\phi_j^{(1)}\}$ from (2.9) are not in the parameter space and ML estimates under the unrestricted AMAR model can be obtained from the EM algorithm (Table 2.5).

Table 2.3: 2×2 Tables with Supplemental Margins for Both Variables

3A:		Y		
		1	2	missing
	1	50	150	30
X	2	75	75	60
	missing	28	60	50

3B:		Y		
		1	2	missing
	1	100	50	30
X	2	75	75	60
	missing	28	60	50

Table 2.4: Estimates of Parameters in the Unrestricted AMAR Model for data in 3A

	Parameter of Interest					Nuisance Parameter			
	θ_{11}	θ_{12}	θ_{21}	θ_{22}	ϕ	$\phi_1^{(0)}$	$\phi_2^{(0)}$	$\phi_1^{(1)}$	$\phi_2^{(1)}$
Non-iterative Estimate:	0.131	0.392	0.239	0.239	0.239	0.130	0.286	0.113	0.636
EM algorithm:	0.131	0.392	0.239	0.239	0.239	0.130	0.286	0.113	0.636

Table 2.5: Estimates of Parameters in the Unrestricted AMAR Model for data in 3B

	Parameters of Interest					Nuisance Parameters			
	θ_{11}	θ_{12}	θ_{21}	θ_{22}	ϕ	$\phi_1^{(0)}$	$\phi_2^{(0)}$	$\phi_1^{(1)}$	$\phi_2^{(1)}$
Non-iterative 'Estimate':	0.308	0.154	0.269	0.269	0.261	0.167	0.286	2.507	-1.476
EM algorithm:	0.297	0.153	0.236	0.314	0.261	0.167	0.286	0.867	0

2.4.2 Examples with $J = 3, K = 2$

In Table 2.6, 6A and 6B give data for the case $J = 3$, $K = 2$. In these cases, the solution set to (2.9) is a straight line and the parameter space for $\{\phi_j^{(1)}\}$ is a unit cube as displayed in Figures 2.1 and 2.2. For the data in 6A, the solution line does not intersect the cube (Figure 2.1), and ML estimates in the unrestricted AMAR model are obtained iteratively (Table 2.7). For the data in 6B, the solution line intersects the cube (Figure 2.2). The non-iterative estimates in table 2.8 obtained from the patterns in which X is observed are the unique ML estimates of $\{\theta_{jk}\}$ in the unrestricted AMAR model, although there are multiple ML estimates for $\{\phi_j^{(1)}\}$.

Table 2.6: 3 × 2 Tables with Supplemental Margins for Both Variables

6A:		Y		
		1	2	missing
	1	100	50	30
X	2	75	75	60
	3	32	67	20
	missing	28	60	50

6B:		Y		
		1	2	missing
	1	50	150	30
X	2	75	75	60
	3	32	67	20
	missing	28	60	50

Figure 2.1: Non-iterative Estimates of $\phi_j^{(1)}$ for data in 6A

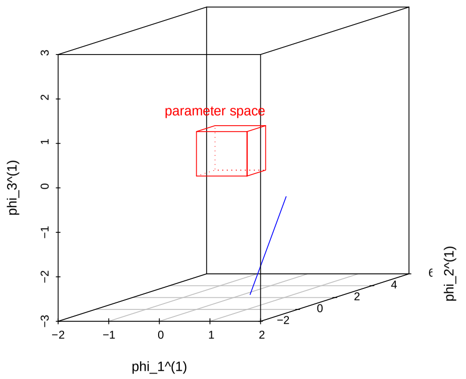

Table 2.7: Estimates of Parameters in the Unrestricted AMAR Model for data in 6A

		Parameters of Interest							Nuisance Parameters				
	θ_{11}	θ_{12}	θ_{21}	θ_{22}	θ_{31}	θ_{32}	ϕ	$\phi_1^{(0)}$	$\phi_2^{(0)}$	$\phi_3^{(0)}$	$\phi_1^{(1)}$	$\phi_2^{(1)}$	$\phi_3^{(1)}$
Non-iterative 'estimate':	0.236	0.118	0.206	0.206	0.076	0.158	0.213	0.167	0.286	0.168	*no solution*		
EM algorithm:	0.235	0.117	0.192	0.219	0.071	0.166	0.213	0.167	0.286	0.168	1	0.037	0

Figure 2.2: Non-iterative Estimates of $\phi_j^{(1)}$ for data in 6B

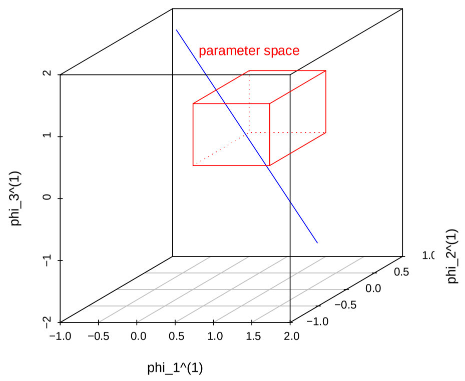

Table 2.8: Estimates of Parameters in the Unrestricted AMAR Model for data in 6B

		Parameter of Interest							Nuisance Parameter				
	θ_{11}	θ_{12}	θ_{21}	θ_{22}	θ_{31}	θ_{32}	ϕ	$\phi_1^{(0)}$	$\phi_2^{(0)}$	$\phi_3^{(0)}$	$\phi_1^{(1)}$	$\phi_2^{(1)}$	$\phi_3^{(1)}$
Non-iterative estimate:	0.103	0.309	0.188	0.188	0.069	0.144	0.198	0.130	0.286	0.168	*multiple solutions*		

2.5 Muscatine Coronary Risk Factor Study

The Muscatine Coronary Risk Factor Study (MCRF) is a longitudinal study of obesity in 4856 school children. Five cohorts (ages 5-7, 7-9, 9-11, 11-13, 13-15) of boys and girls were measured for height and weight between 1977 and 1981. Children with relative weight greater than 110 percent of the median weight for their age-gender-height group were classified as obese and at any time point about 20 percent of the children were obese. We are interested in estimating obesity rates over time and evaluating whether or not these rates differ by gender. The study was first presented by Woolson and Clarke (1984), and further analyses can be found in, e.g., Baker (1995), Ekholm and Skinner (1998), Lipsitz, Parzen and Molenberghs (1998) and Birmingham and Fitzmaurice (2002).

The analysis is complicated by the study design. Both cross-sectional and longitudinal information about age trends in obesity rates were presented in the data. Due to cohort effects, cross-sectional age trends in obesity rates may be different from longitudinal trends. Ekholm and Skinner (1998) found no statistical cohort effects. Therefore, in our analyses, cohort effects are assumed negligible and data are pooled across five age-group cohorts. The data for 1977 and 1981 are given in table 2.9.

Table 2.9: Tables of data from Muscatine Coronary Risk Factor Study

girls:		**1981**		
		1	2	missing
1977	1[1]	701	98	497
	2	59	111	183
	missing	408	139	174

boys:		**1981**		
		1	2	missing
1977	1	699	98	566
	2	72	116	141
	missing	473	125	196

The analysis is further complicated by the substantial non-response. Only 40 percent of children provided complete records in 1977 and 1981. In addition to the complete records, there are three non-response patterns, specifically, two patterns

[1] 1 = not obese, 2 = obese.

with one missing response and one pattern with two missing responses. Baker (1995) reported there were two main reasons for non-response: (1) no parental consent form was received and (2) the child was not in school on the examination day. For girls, the missingness of obese status in 1981 is found to depend on the missingness in 1977 using a Chi-square test (p-value < 0.0001). Furthermore, girls measured and classified as obese in 1977 were more likely to have missing data in 1981 than those classified as non-obese (p-value < 0.0001 based on a chi-square test). The estimates of girls' obesity rates and missing probabilities in the AMAR model discussed above are presented in table 2.10. For the unrestricted AMAR model, the estimate from (2.9) of $\{\phi_1^{(1)}, \phi_2^{(1)}\}$ is $(0.274, 0.121)$ which is in the parameter space, so closed form estimates of the parameters are available. A bootstrap approach was used to estimate standard errors. If a bootstrap sample leads to 'estimates' of $\{\phi_j^{(1)}\}$ that lie outside of the parameter space, the EM algorithm is used to obtain the ML estimates. Among the 1000 bootstrap samples, 23.2% samples yield 'estimates' of $\{\phi_j^{(1)}\}$ that are outside of the parameter space.

Likelihood ratio tests can be utilized to test the two submodels discussed above against the more general unrestricted AMAR model. Denote the unrestricted AMAR model as M1, the restricted AMAR model as M2 and the restricted MAR model in section 2.2 as M3, and let l_{max} present maximum loglikelihood. We find that $-2(l_{max}(M2) - l_{max}(M1)) = -2(-4569.823 + 4535.292) = 69.062$, which yields a p-value < 0.0001 when compared to χ_2^2. There is strong evidence that the restricted AMAR model does not fit the data. On the other hand, $l_{max}(M3)$ is close to $l_{max}(M1)$, we can't differentiate the restricted MAR model from the unrestricted AMAR model.

Similarly for the boys, the estimate from (2.9) of $\{\phi_1^{(1)}, \phi_2^{(1)}\}$ in the unrestricted AMAR model is $(0.228, 0.325)$ which is in the parameter space, closed form estimates of the parameters are available. Among 1000 bootstrap samples, there are 2.8% samples with $\{\phi_j^{(1)}\}$ outside of the parameter space. For the likelihood ratio tests, $-2(l_{max}(M2) - l_{max}(M1)) = -2(-4748.480 + 4713.027) = 70.906 > \chi_2^2(0.05)$, the restricted AMAR model does not fit the data. While $l_{max}(M3)$ is close to $l_{max}(M1)$, we can't differentiate the restricted MAR model from the unrestricted AMAR model (Table 2.11).

Note that, for the boys, $\hat{\phi}_1^{(0)}$ and $\hat{\phi}_2^{(0)}$ are nearly the same which suggests a MCAR mechanism. For girls, however, these estimates are quite different, suggesting the fact that girls measured and classified as obese in 1977 are less likely to be present and measured in 1981 than those measured and classified as non-obese in 1977. This is also noted by Ekholm and Skinner (1998).

2.6 Discussion

The main goal of the current chapter is to illustrate a non-MAR model that incorporates features that we tend to associate with "randomly missing" data. This "almost MAR" model is considered in the case of bivariate categorical data, and it is shown that ML estimates have interesting features, including the discarding of data that would at first glance appear to contain information about the parameters of interest.

There appears to be very little existing literature on missing data mechanisms of the type considered here; most of the work on NMAR mechanisms concerns the situation where missingness depends directly on outcomes of interest, or on latent variables such as the slope of a repeatedly measured variable (e.g. Little and Rubin 2002, chapter 15). Perhaps the closest work to that presented here concerns the "latent ignorable" missing data mechanisms proposed to model missing data in the presence of noncompliance with a treatment (Frangakis and Rubin, 1999; Peng, Little and Raghunathan, 2004). In these cases, there is a binary compliance variable that indicates whether an individual would comply with a treatment if assigned to it. In a clinical trial, this indicator is fully observed for cases in the active treatment group, but is completely missing for cases in the control group, since those cases do not have an access to the active treatment. The latent ignorable model assumes MAR within subpopulations defined by the compliance indicator.

Some extensions of the ideas discussed here include the following:

(A) Models for bivariate data involving continuous or ordinal variables, with the same pattern and mechanism as that described here.

(B) The additional of fully observed covariates to the data structure considered here. The latent ignorable mechanism for missing data when there is noncompliance with a treatment is a special case of this structure.

(C) Extensions to two sets of variables, where some variables can be assumed to be MCAR, and other variables would be MAR if variables in the other set were fully observed.

(D) extensions of (C) to more than two blocks of variables; a variety of extensions seem possible.

We plan to pursue these extensions in future work.

Table 2.10: Estimates of Girls' Obesity Rates

		Obesity Rate				Nuisance Parameter				Observed Data
	θ_{11}	θ_{12}	θ_{21}	θ_{22}	ϕ	$\phi_1^{(0)}$	$\phi_2^{(0)}$	$\phi_1^{(1)}$	$\phi_2^{(1)}$	Loglikelihood
complete-case estimate	0.723	0.101	0.061	0.115	—	—	—	—	—	—
	(0.014)	(0.010)	(0.008)	(0.010)						
restricted MAR								$\phi_1^{(1)} = \phi_2^{(1)}$		
					ϕ	$\phi_1^{(0)}$	$\phi_2^{(0)}$	$\phi^{(1)}$		
	0.685	0.099	0.073	0.143	0.304	0.383	0.518	0.241		
	(0.012)	(0.009)	(0.009)	(0.010)	(0.010)	(0.006)	(0.023)	(0.016)		-4535.605
restricted AMAR						$\phi_j^{(0)} = \phi_j^{(1)}, j = 1, 2$				
					ϕ	ϕ_1		ϕ_2		
	0.683	0.103	0.070	0.143	0.304	0.335		0.455		
	(0.011)	(0.009)	(0.008)	(0.010)	(0.010)	(0.006)		(0.022)		-4569.823
unrestricted AMAR					ϕ	$\phi_1^{(0)}$	$\phi_2^{(0)}$	$\phi_1^{(1)}$	$\phi_2^{(1)}$	
	0.690	0.096	0.074	0.140	0.304	0.383	0.518	0.274	0.121	
	(0.012)	(0.010)	(0.009)	(0.010)	(0.010)	(0.006)	(0.023)	(0.034)	(0.122)	-4535.292

Table 2.11: Estimates of Boys' Obesity Rates

		Obesity Rate				Nuisance Parameter				Observed Data
	θ_{11}	θ_{12}	θ_{21}	θ_{22}	ϕ	$\phi_1^{(0)}$	$\phi_2^{(0)}$	$\phi_1^{(1)}$	$\phi_2^{(1)}$	Loglikelihood
complete-case estimate	0.710	0.099	0.073	0.118	—	—	—	—	—	—
	(0.015)	(0.010)	(0.008)	(0.010)						
restricted MAR								$\phi_1^{(1)} = \phi_2^{(1)}$		
					ϕ	$\phi_1^{(0)}$	$\phi_2^{(0)}$	$\phi^{(1)}$		
	0.709	0.097	0.075	0.118	0.319	0.415	0.429	0.247		
	(0.011)	(0.009)	(0.008)	(0.008)	(0.009)	(0.006)	(0.025)	(0.015)		-4713.142
restricted AMAR						$\phi_j^{(0)} = \phi_j^{(1)}, j = 1, 2$				
					ϕ	ϕ_1		ϕ_2		
	0.709	0.098	0.075	0.118	0.319	0.360		0.375		
	(0.011)	(0.009)	(0.008)	(0.008)	(0.009)	(0.005)		(0.023)		-4748.480
unrestricted AMAR					ϕ	$\phi_1^{(0)}$	$\phi_2^{(0)}$	$\phi_1^{(1)}$	$\phi_2^{(1)}$	
	0.707	0.099	0.074	0.120	0.319	0.415	0.429	0.228	0.325	
	(0.013)	(0.009)	(0.008)	(0.009)	(0.009)	(0.006)	(0.025)	(0.037)	(0.153)	-4713.027

CHAPTER III

Estimating Treatment Effects in Randomized Clinical Trials with Non-compliance and Missing Outcomes

Abstract We analyze randomized trials with active treatment verses control treatment, where treatments are subject to all-or-none compliance and outcomes have missing values. In addition to latent ignorability (Frangakis and Rubin, 1999), we further specify two assumptions for principal compliance and two assumptions for missing outcome to identify the model. In each of four scenarios defined by combinations of these assumptions, we derive maximum likelihood (ML) estimates by using the EM algorithm, as well as non-iterative ML estimates by implementing pattern-mixture models with covariates (Little and Wang, 1996). This shows that, under certain conditions, the method-of-moments (MOM) estimates are ML estimates. We show that the models of principal compliance determine which type of analysis is used to estimate treatment efficacy, per-protocol analysis or IV estimation with the treatment assignment indicator as the instrumental variable. On the other hand, we show that the assumptions for missing outcome determine whether MOM estimates are ML estimates or not. We apply our methods to data from a double-blinded randomized clinical trials with clozapine vs. haloperidol for patients with refractory schizophrenia.

keywords: causal inference; non-compliance; non-response; per-protocol analysis; instrumental variables; maximum likelihood; EM algorithm.

3.1 Introduction

Analysis of randomized clinical trials involving human subjects is often complicated by non-compliance to treatment assignments and missing outcomes. Compliance is often associated with the effects of treatments, and may vary according to participant characteristics. For example, due to severe side effects in a clinical trial to test the efficacy of a new drug, some participants may not comply with the assignment of new drug and will switch to take the conventional drug. In psychiatric trials, subjects' mental health condition may affect their willingness or ability to comply with study protocols. We consider a clinical trial involving random assignments to an active treatment or a control treatment and assume the active treatment is subject to all-or-none compliance. This is in contrast to an alternative situation of partial compliance (Baker 1997).

Ignoring non-compliance information, a standard intention-to-treat (ITT) analysis compares the difference in outcome distributions based on treatment assignments. It provides a valid measurement of treatment effectiveness (the effect of the treatment assignment), but is potentially biased for estimating treatment efficacy (the effect of the treatment itself), which is more often the main interest. By comparing the difference in outcome distributions between treatments actually received, the as-treated (AT) analysis attempts to directly estimate the effect of the treatment itself, but is subject to selection bias since the randomization is not preserved. A more recent approach to estimating treatment efficacy in randomized trials with non-compliance is to treat the randomization as an instrumental variable (IV), in economic parlance. Based on certain assumptions on outcomes for non-compliers in both treatment groups, the IV estimator corrects the ITT estimator for noncompliance. That is, it estimates the treatment effect among the compliers. This approach maintains the properties of randomization and yields a direct estimate of treatment efficacy.

Principal compliance is a special case of principal stratification (Frangakis and Rubin 2002), where individuals are stratified according to the values of the post-treatment variable under both treatments, rather than simply under the treatment actually assigned, since the post-treatment variable is compliance. Principal compliance differs from observed compliance. Participants are said to be never-takers if they take control treatment no matter which treatment they are assigned, and participants are compliers if they comply with their assignments. In randomized clinical trials we consider here, participants in the active treatment group may switch to take the control treatment, while those in the control treatment group don't have access to the active treatment. Participants in the control group take their assignments and therefore are observed compliers, however whether they are never-takers or principal compliers is unknown. Besides missing principal compliance for those in the control group, analyzing randomized clinical trials may be further complicated with missing outcomes due to loss to follow-up or non-response. Some methods have recently been developed for an IV estimator by accounting for non-compliance and subsequent non-responses in clinical trials. Under assumptions of latent ignorability (outcomes are missing at random conditional on latent compliance status and treatment assignments) and compound exclusion restrictions (the missingness and potential values of outcomes are independent of treatment assignments for never-takers), Frangakis and Rubin (1999) proposed a method-of-moments (MOM) IV estimator which is asymptotically valid. Under the same assumptions, Zhou and Li (2006) developed maximum likelihood (ML) estimates when the outcome is binary and O'Malley and Normand (2005) obtained ML estimators for normal distributed outcomes using an EM algorithm.

Another approach to estimating treatment efficacy is per-protocol (PP) analysis, which restricts analysis to observed compliers. By classifying participants according to treatments they actually received, PP analysis is subject to selection bias, since participants observed complying with a treatment may be a biased sample of participants randomized to that treatment. Although the bias may be reduced by adjusting for covariates, it remains a concern. In practice, to compare treatments in randomized clinical trials, a choice has to be made among AT, IV and PP analyses. Little, Long and Lin (2008) compared these analyses when there is noncompliance in clinical trials under various assumptions and examined their design implications.

In this chapter, we consider two-arm clinical trials with a categorical outcome subject to both noncompliance and missing outcome. To identify the model, we specify two assumptions for principal compliance, ER (exclusion restriction which indicates the outcome is independent of treatment assignments for never-takers) or NCEC (none compliance effect in controls which implies the distribution of the outcome is same for principal compliers and never-takers in the control group), and two assumptions for missing outcome, ER (there is no effect of treatment assignments on missingness of the outcome for never-takers) or NCEC (there is no effect of principal compliance on missingness of the outcome for participants in the control group). In each of four scenarios defined by combinations of these assumptions, we derive ML estimates by using the EM algorithm, as well as non-iterative ML estimates by implementing pattern-mixture models with covariates (Little and Wang, 1996), which shows that, under certain conditions, the MOM estimates are ML estimates. We find the assumptions of principal compliance determine which type of analysis is used to estimate treatment efficacy, PP analysis or IV estimation, whereas the assumptions of missing outcome decide whether MOM estimates are ML estimates or not. Both complier average causal effect (CACE) or ITT estimands can be viewed as outputs from these models.

The chapter is organized as follows. Section 3.2 presents notations and assumptions. The MOM estimator and likelihood-based methodology are developed in section 3.3 where ER is assumed for missing outcome. EM algorithm based on the selection model and closed-form estimators based on the pattern-mixture model are both given there. Section 3.4 gives the MOM estimator and likelihood-based methodology where NCEC is assumed for missing outcome. In section 3.5, applications of our proposed methods are illustrated in a real study. Conclusions are made and future work are given in section 3.6. Some technical details of our methods are provided in appendix.

3.2 Notation and Assumptions

Notation: we consider a randomized trial where participants are assigned to either active treatment ($T = 1$) or control ($T = 0$). C represents the participant's principal compliance, with $C = 1$ denoting compliers and $C = 0$ denoting never-takers. Participants in the active treatment group may comply with their treatment assignments ($C = 1$) or switch to take the control treatment ($C = 0$). The compliance status of those in the control group is unknown. We consider an outcome Y with K levels. Due to loss to follow-up or dropout, some participants have missing values (Table 3.1). Let M^Y denotes the missing-data indicator for Y such that M^Y is 1 if Y is missing and M^Y is 0 if Y is observed. It is worth noting that treatment assignment T implies the missingness of compliance status C. For participants in active treatment group ($T = 1$), we know their compliance status, while for those in control group ($T = 0$), the compliance status is missing. Let M^C denote the missing-data indicator for compliance status C with 1 and 0 representing missing and observed respectively, then we have $M^C = 1 - T$.

Table 3.1: Randomized Clinical Trials with Non-compliance and Non-responses

T	C	Y	
1	1	■	$n_{(0),1k}$
		?	$n_{(1),1+}$
	0	■	$n_{(0),0k}$
		?	$n_{(1),0+}$
0	?	■	$n_{(2),+k}$
		?	$n_{(3),++}$

According to whether C and Y are observed or missing, there are totally four missing-data patterns. let P_r denote the set of sample cases with pattern type r, $r = 0, \ldots, 3$. Let $n_{(0),jk}$ be the count of complete cases with $C = j, Y = k$, $n_{(1),j+}$ be the count of cases with $C = j$ and Y missing, $n_{(2),+k}$ be the count of cases with $Y = k$ and C missing, and $n_{(3),++}$ be the count of cases with both C and Y missing. Then $n_0 = \sum_{j=0}^{1} \sum_{k=0}^{K-1} n_{(0),jk}$, $n_1 = \sum_{j=0}^{1} n_{(1),j+}$, $n_2 = \sum_{k=0}^{K-1} n_{(2),+k}$, $n_3 = n_{(3),++}$ are the number of cases in each pattern and $n = \sum_{r=0}^{3} n_r$ is the total sample size.

Parameters of Interest: the parameters of interest are $\theta = \{\theta_{jk}^{(t)}\}$, where $\theta_{jk}^{(t)} = P(C = j, Y = k|T = t)$ with $\sum_{j=0}^{1} \sum_{k=0}^{K-1} \theta_{jk}^{(t)} = 1$. Treatment efficacy can be measured in functions of θ.

Missing-data Mechanism: with compliance status and outcome having missing values, a missing-data mechanism has to be specified to make valid inferences. As we state before, the missingness of C is decided by treatment assignment T:

$$P(M^C = 1|T, C, Y) = \begin{cases} 0 & T = 1 \\ 1 & T = 0 \end{cases}$$

For missingness of the outcome Y, we assume latent ignorability as defined in Frangakis and Rubin (1999). Thus, we assume that

$$P(M^Y = 1|T = t, C = j, Y = k) = \phi_j^{(t)} \tag{3.1}$$

where $t = 0, 1;\ j = 0, 1;\ k = 0, 1, \ldots, K-1$. Given the fact that $M^C = 1 - T$, the missingness of Y depends on M^C and C, but not on the value of Y, which is exactly the same as AMAR defined in chapter 2. Thus, the missing-data mechanism for Y is ignorable conditional on the latent variable C.

For the data in the active treatment group, the degrees of freedom are $2K+1$, which comprise K for complete cases with $C = 1$ plus 1 for those with missing Y, plus K for complete cases with $C = 0$ plus 1 for those with missing Y, minus 1 for the total which is considered fixed at $T = 1$. For the data in the control treatment group, the degrees of freedom are K, which comprise K for supplemental margin on Y, plus 1 for cases with missing C and Y, minus 1 for the total which is considered fixed at $T = 0$. Therefore, the total degrees of freedom in the data are $3K+1$, which is less than the total number of parameters $4K+2$. To identify the modeling, we specify a number of further assumptions, and later the results of assuming selecting various subsets of these to hold:

Assumption 1: Stable unit treatment value assumption (SUTVA) (Rubin, 1978), which implies that compliance and potential outcomes for individuals are not affected by treatment assignments and outcomes of other individuals in the sample.

Assumption 2: Randomization. As an attribute of participants, principal compliance is viewed as being determined before assignments of treatments and it is a covariate where value is fixed at baseline. By the property of randomization, compliance is independent of treatment assignment. This implies that C has the same distribution in the active treatment group as in the control treatment group.

$$P(C = j|T = 1) = P(C = j|T = 0) \tag{3.2}$$

Assumption 3: ER for Y. Never-takers actually receive the control treatment no matter which treatment they are assigned. According to this assumption, there is no effect of the treatment assignment on the distribution of their outcome Y. That is,

$$P(Y = k|T = 1, C = 0) = P(Y = k|T = 0, C = 0) \tag{3.3}$$

Assumption 3 is closely related to exclusion restriction assumptions in the traditional instrumental variable approach (Durbin, 1954; Goldberger, 1972; Angrist et al, 1996) and biomedical applications (Baker and Lindeman, 1994, Sommer and Zeger, 1991). This assumption may not hold in all studies, for example, in an unblinded study where the failure to complying with an assigned treatment may have a lingering psychological effect on the patient that affects the outcome. In a double-blind trial where the treatment assignment is not known for both the patient and the physician, the exclusion restriction for the outcome is reasonable.

Assumption 4: NCEC for Y. The distribution of Y is assumed to be same for principal compliers and never-takers in the control group.

$$P(Y = k|C = 1, T = 0) = P(Y = k|C = 0, T = 0) \tag{3.4}$$

This is a strong assumption and widely viewed as unacceptable, since in the control group, principal compliers and never-takers may have different characteristics that are related to the outcome. White (2005) argues that NCEC may be plausible in double-blinded prevention trials if noncompliance relates to treatment discontinuation and the active agent has low rates of adverse events. Although NCEC can be weakened by adjusting for covariates, it remains a questionable assumption that needs to be carefully examined.

Assumption 5: ER for M^Y. Similar to Assumption 3, this specifies that there is no effect of randomized treatment assignments on M^Y for never-takers. Thus,

$$P(M^Y = 1|T = 1, C = 0) = P(M^Y = 1|T = 0, C = 0) \tag{3.5}$$

This assumption is stronger than Assumption 3. Never-takers always end up taking the control treatment, however, their treatment assignments may be more likely to affect their missingness. Although it is impossible to test this assumption, we can develop a sensitivity analysis to evaluate its influence on the estimators of treatment efficacy. We will use the term "compound ER" for the combined Assumptions 3 and 5.

Assumption 6: NCEC for M^Y. Similar to Assumption 4, this specifies that there is no effect of principal compliance on M^Y for participants in the control group. That is,

$$P(M^Y = 1|C = 1, T = 0) = P(M^Y = 1|C = 0, T = 0) \tag{3.6}$$

This is also a strong and questionable assumption that needs to be critically evaluated, since the compliance status of individuals in the control group may affect their missingness. A sensitivity analysis can be developed to evaluate the influence of this assumption on the estimators of treatment efficacy. We will use the term "compound ER" for the combined Assumptions 4 and 6.

If, in addition to latent ignorability, Assumption 5 is also assumed, Y is still not missing at random (NMAR) since M^Y depends on C, which is missing in the control group. But if Assumption 6 is specified instead, Y will then be missing at random (MAR), since M^Y is independent of the missing C in the control group.

To estimate CACE, $\delta = E(Y|T = 1, C = 1) - E(Y|T = 0, C = 1) = \mu_{T=1,C=1} - \mu_{T=0,C=1}$, under Assumptions 1 and 2, we consider all four combinations of assumptions for Y (ER or NCEC) and M^Y (ER or NCEC). In section 3, MOM and ML estimators are obtained when Assumption 5 is added to the assumption of latent ignorability. In section 4, these estimators are obtained when the missing-data mechanism of Y is MAR as the additional Assumption 6 is applied to M^Y.

3.3 Estimation when ER is assumed for M^Y

3.3.1 ER is assumed for Y

MOM estimator. Frangakis and Rubin (1999) proposed an estimator for CACE using the observed data. By Assumption 2, the probability of compliance $P(C = 1)$ is estimated by the proportion of participants in the active treatment group who take the active treatment. Under latent ignorability, participants with Y observed are a random sample of those with same treatment T and compliance C, therefore, the mean of outcomes for principal compliers in the active treatment group $\mu_{T=1,\,C=1}$ can be estimated using the mean of observed outcomes for those in the active treatment group who take their assignments. In summary,

$$\hat{P}(C = 1) = \frac{n_{(0),1+} + n_{(1),1+}}{n_0 + n_1} \tag{3.7}$$

$$\hat{\mu}_{T=1,\,C=1} = \sum_{k=0}^{K-1} k \frac{n_{(0),1k}}{n_{(0),1+}} \tag{3.8}$$

Since the principal compliance status for participants in the control group is unknown, we can apply the observed information in the control group and those for never-takers in the active treatment group to represent $\mu_{T=0,\,C=1}$, the mean of outcomes for compliers in the control group. By Bayes' theorem and Assumptions 2, 3 and 5, we have:

$$\begin{aligned}&\mu_{T=0,\,C=1}\\&= \frac{\mu_{T=0,\,M^Y=0}P(M^Y = 0|T = 0) - \mu_{T=1,\,C=0}P(M^Y = 0|T = 1, C = 0)P(C = 0)}{P(M^Y = 0|T = 0) - P(M^Y = 0|T = 1,\, C = 0)P(C = 0)}\end{aligned} \tag{3.9}$$

which subtract never-taker outcomes from the overall observed outcomes in the control group, while never-taker outcomes are represented by outcomes of never-takers in the treated group. Thus (3.9) can be directly estimated from the observed data:

$$
\begin{aligned}
&\hat{\mu}_{T=0,\,C=1} \\
&= \frac{\hat{\mu}_{T=0,\,M^Y=0}\hat{P}(M^Y=0|T=0) - \hat{\mu}_{T=1,\,C=0}\hat{P}(M^Y=0|T=1,C=0)\hat{P}(C=0)}{\hat{P}(M^Y=0|T=0) - \hat{P}(M^Y=0|T=1,C=0)\hat{P}(C=0)} \\
&= \sum_{k=0}^{K-1} k \frac{n_{(2),+k}(n_0+n_1) - n_{(0),\,0k}(n_2+n_3)}{n_2(n_0+n_1) - n_{(0),\,0+}(n_2+n_3)} \qquad (3.10)
\end{aligned}
$$

By using estimators in (3.8) and (3.10), we then can estimate CACE as the following:

$$
\begin{aligned}
\hat{\delta}_{IV} &= \hat{\mu}_{T=1,\,C=1} - \hat{\mu}_{T=0,\,C=1} \\
&= \sum_{0}^{K-1} k \cdot (\frac{n_{(0),1k}}{n_{(0),1+}} - \frac{n_{(2),+k}(n_0+n_1) - n_{(0),\,0k}(n_2+n_3)}{n_2(n_0+n_1) - n_{(0),\,0+}(n_2+n_3)}) \qquad (3.11)
\end{aligned}
$$

which is sometimes termed the IV estimator, since the randomization indicator is used as the instrumental variable.

EM algorithm. The likelihood for our modeling has the form:

$$
\begin{aligned}
&L(\theta,\phi|T,C_{obs},Y_{obs},M^C,M^Y) \\
&= \sum_{C_{mis}}\sum_{Y_{mis}}\left\{\prod_{i=1}^{n} p(C_i,Y_i|T_i,\theta)p(M_i^C|T_i,C_i,Y_i)\right. \\
&\qquad\qquad \left. p(M_i^Y|M_i^C,T_i,C_i,Y_i,\phi_j^{(0)},\phi_j^{(1)})\right\} \\
&= \sum_{C_{mis}}\sum_{Y_{mis}}\left\{\prod_{i=1}^{n}\prod_{j=0}^{1}\prod_{k=0}^{K-1} {\theta_{jk}^{(1)}}^{I(C_i=j,Y_i=k|T_i=1)} {\theta_{jk}^{(0)}}^{I(C_i=j,Y_i=k|T_i=0)}\right. \\
&\qquad\qquad \prod_{j=0}^{1} {\phi_j^{(1)}}^{I(T_i=1,C_i=j,M_i^Y=1)} (1-\phi_j^{(1)})^{I(T_i=1,C_i=j,M_i^Y=0)} \\
&\qquad\qquad \left.\prod_{j=0}^{1} {\phi_j^{(0)}}^{I(T_i=0,C_i=j,M_i^Y=1)} (1-\phi_j^{(0)})^{I(T_i=0,C_i=j,M_i^Y=0)}\right\}.
\end{aligned}
$$

where $I(.)$ is the indicator function and $\phi = \{\phi_j^{(1)}, \phi_j^{(0)}\}$.

To define the E step of EM, let $({\theta_{jk}^{(1)}}^{(t)}, {\theta_{jk}^{(0)}}^{(t)}, {\phi_j^{(1)}}^{(t)}, {\phi_j^{(0)}}^{(t)})$ denote the parameter estimates at iteration t, and $n_{(r),jk}^{(t)}$ be the estimate of cell frequency for $C=j, Y=k$

in pattern P_r. The E step distributes the partially classified observations into the table according to the corresponding probabilities:

$$\begin{aligned}
n_{(1),jk}^{(t)} &= n_{(1),j+} \cdot \frac{\theta_{jk}^{(1)^{(t)}}}{\theta_{j+}^{(1)^{(t)}}} \ , \\
n_{(2),jk}^{(t)} &= n_{(2),+k} \cdot \frac{(1-\phi_j^{(0)^{(t)}})\theta_{jk}^{(0)^{(t)}}}{\sum_{j=0}^{1}(1-\phi_j^{(0)^{(t)}})\theta_{jk}^{(0)^{(t)}}} \ , \\
n_{(3),jk}^{(t)} &= n_{(3),++} \cdot \frac{\phi_j^{(0)^{(t)}}\theta_{jk}^{(0)^{(t)}}}{\sum_{j=0}^{1}\phi_j^{(0)^{(t)}}\theta_{j+}^{(0)^{(t)}}} \ .
\end{aligned}$$

Under the constraints $\theta_{0k}^{(1)} = \theta_{0k}^{(0)}$, $k = 0, \ldots, K-1$ implied by Assumptions 2 and 3 and $\phi_0^{(1)} = \phi_0^{(0)}$ implied by Assumption 5, the M step calculates new parameter estimates as:

$$\begin{aligned}
\theta_{0k}^{(0)^{(t+1)}} &= \theta_{0k}^{(1)^{(t+1)}} = \frac{n_{(0),0k} + n_{(1),0k}^{(t)} + n_{(2),0k}^{(t)} + n_{(3),0k}^{(t)}}{n} \ , \\
\theta_{1k}^{(0)^{(t+1)}} &= \frac{n_{(2),1k}^{(t)} + n_{(3),1k}^{(t)}}{n_{(2),1+}^{(t)} + n_{(3),1+}^{(t)}} \cdot \frac{n_{(0),1+} + n_{(1),1+} + n_{(2),1+}^{(t)} + n_{(3),1+}^{(t)}}{n} \ , \\
\theta_{1k}^{(1)^{(t+1)}} &= \frac{n_{(0),1k} + n_{(1),1k}^{(t)}}{n_{(0),1+} + n_{(1),1+}} \cdot \frac{n_{(0),1+} + n_{(1),1+} + n_{(2),1+}^{(t)} + n_{(3),1+}^{(t)}}{n} \ , \\
\phi_0^{(1)^{(t+1)}} &= \phi_0^{(0)^{(t+1)}} = \frac{n_{(1),0+} + n_{(3),0+}^{(t)}}{n_{(0),0+} + n_{(1),0+} + n_{(2),0+}^{(t)} + n_{(3),0+}^{(t)}} \ , \\
\phi_1^{(1)} &= \frac{n_{(1),1+}}{n_{(0),1+} + n_{(1),1+}} \ , \\
\phi_1^{(0)^{(t+1)}} &= \frac{n_{(3),1+}^{(t)}}{n_{(2),1+}^{(t)} + n_{(3),1+}^{(t)}} \ .
\end{aligned}$$

The E-step and M-step alternate until the parameter estimates converge.

Note that $\phi_1^{(1)}$ is estimated directly and is unchanged throughout the EM algorithm. Complete-case estimates can be chosen as the starting values of $\{\theta_{jk}^{(t)}\}$ and frequency estimates of $\{\phi_j^{(1)}\}$ in the data can be taken as initial values of $\{\phi_j^{(t)}\}$.

As a result, CACE can be estimated by:

$$\begin{aligned}\hat{\delta}_{IV} &= \hat{\mu}_{T=1,C=1} - \hat{\mu}_{T=0,C=1} \\ &= \sum_{0}^{K-1} k \cdot \left(\frac{\hat{\theta}^{(1)}_{1k}}{\hat{\theta}^{(1)}_{1+}} - \frac{\hat{\theta}^{(0)}_{1k}}{\hat{\theta}^{(0)}_{1+}}\right)\end{aligned}$$

Non-iterative ML estimates. After applying $K+1$ constraints implied by Assumption 2 and compound ER in Assumptions 3 and 5, $\theta^{(1)}_{0k} = \theta^{(0)}_{0k}$, $k = 0, \ldots, K-1$ and $\phi^{(1)}_0 = \phi^{(0)}_0$, our model can be exactly identified. So non-iterative ML estimates may exist. To find non-iterative ML estimates, the likelihood function can be factorized into the pattern-mixture components. The parameters corresponding to the pattern-mixture model can be defined as:

$$\begin{aligned}&\alpha_{(0),jk} = P(C = j, Y = k | T = 1, M^C = M^Y = 0), \\ &\alpha_{(1),j+} = P(C = j | T = 1, M^C = 0, M^Y = 1), \\ &\alpha_{(2),+k} = P(Y = k | T = 0, M^C = 1, M^Y = 0), \\ &\pi_0 = P(M^X = 0, M^Y = 0 | T = 1),\ \ \pi_1 = P(M^X = 0, M^Y = 1 | T = 1), \\ &\pi_2 = P(M^X = 1, M^Y = 0 | T = 0),\ \ \pi_3 = P(M^X = 1, M^Y = 1 | T = 0).\end{aligned} \tag{3.12}$$

where $0 \leq j \leq 1$, $0 \leq k \leq K-1$ and the following constraints apply:

$$\sum_{j=0}^{1}\sum_{k=0}^{K-1} \alpha_{(0),jk} = 1,\ \sum_{j=0}^{1} \alpha_{(1),j+} = 1,\ \sum_{k=0}^{K-1} \alpha_{(2),+k} = 1,\ \sum_{r=0}^{3} \pi_r = 1.$$

These parameters can be expressed in terms of (θ, ϕ) as follows:

$$\begin{aligned}\alpha_{(0),jk} &= \frac{(1-\phi^{(1)}_j)\theta^{(1)}_{jk}}{\sum_{j=0}^{1}(1-\phi^{(1)}_j)\theta^{(1)}_{j+}} \\ \alpha_{(1),j+} &= \frac{\phi^{(1)}_j \theta^{(1)}_{j+}}{\sum_{j=0}^{1}\phi^{(1)}_j \theta^{(1)}_{j+}} \\ \alpha_{(2),+k} &= \frac{\sum_{j=0}^{1}(1-\phi^{(0)}_j)\theta^{(0)}_{jk}}{\sum_{j=0}^{1}(1-\phi^{(0)}_j)\theta^{(0)}_{j+}}\end{aligned}$$

$$\begin{aligned}\pi_0 &= \sum_{j=0}^{1}(1-\phi_j^{(1)})\theta_{j+}^{(1)},\ \pi_1 = 1-\pi_0\\ \pi_2 &= \sum_{j=0}^{1}(1-\phi_j^{(0)})\theta_{j+}^{(0)},\ \pi_3 = 1-\pi_2 \qquad (3.13)\end{aligned}$$

Letting (α, π) represent the parameters in (3.12), then the likelihood can be written as

$$\begin{aligned}&L(\alpha, \pi | T, C_{obs}, Y_{obs}, M^C, M^Y)\\ &= \prod_{i=1}^{n} p(M_i^C, M_i^Y | T_i) \prod_{i \in p_0} p(C_i, Y_i | T_i = 1, M_i^C = 0, M_i^Y = 0)\\ &\times \prod_{i \in p_1} p(C_i | T_i = 1, M_i^C = 0, M_i^Y = 1)\\ &\times \prod_{i \in p_2} p(Y_i | T_i = 0, M_i^C = 1, M_i^Y = 0)\\ &= \prod_{r=0}^{3} \pi_r^{n_r} \prod_{j=0}^{1}\prod_{k=1}^{K} \alpha_{(0),jk}^{n_{(0),jk}} \prod_{j=0}^{1} \alpha_{(1),j+}^{n_{(1),j+}} \prod_{k=0}^{K-1} \alpha_{(2),+k}^{n_{(2),+k}}\ . \qquad (3.14)\end{aligned}$$

After substituting the transformations in (3.13), we can find closed-form ML estimators for (θ, ϕ) by maximizing (3.14) under the constraints $\theta_{0k}^{(1)} = \theta_{0k}^{(0)}$, $k = 0, \ldots, K-1$ and $\phi_0^{(1)} = \phi_0^{(0)}$. We obtain:

$$\begin{aligned}\hat{\theta}_{jk}^{(1)} &= \frac{n_{(0),jk}}{n_{(0),j+}} \cdot \frac{n_{(0),j+} + n_{(1),j+}}{n_0 + n_1}\ ,\\ \hat{\theta}_{0k}^{(0)} &= \hat{\theta}_{0k}^{(1)},\\ \hat{\theta}_{1k}^{(0)} &= \frac{(n_0+n_1)n_{(2),+k} - (n_2+n_3)n_{(0),0k}}{(n_0+n_1)n_2 - (n_2+n_3)n_{(0),0+}} \cdot \frac{n_{(0),1+} + n_{(1),1+}}{n_0+n_1}\ ,\\ \hat{\phi}_j^{(1)} &= \frac{n_{(1),j+}}{n_{(1),j+} + n_{(0),j+}}\ ,\\ \hat{\phi}_0^{(0)} &= \hat{\phi}_0^{(1)},\\ \hat{\phi}_1^{(0)} &= \frac{\frac{n_3}{n_2+n_3} - \frac{n_{(1),0+}}{n_0+n_1}}{\frac{n_{(0),1+}+n_{(1),1+}}{n_0+n_1}}\ .\end{aligned}$$

This approach yields ML estimates, providing the estimates lie within the parameter space, that is the probabilities lie between zero and one. The expressions for $\{\hat{\theta}^{(t)}_{jk}, \hat{\phi}^{(1)}_j, \hat{\phi}^{(0)}_0\}$ always yield estimates in $[0,1]$. The estimate $\hat{\phi}^{(0)}_1$ given above, however, may or may not fall in $[0,1]$. If it does, then estimates from this approach are ML estimates. If not, this approach fails to yield ML estimates of the parameters of interest. In this case, the EM algorithm can still be used. Furthermore, when $\hat{\phi}^{(0)}_1 \in [0,1]$, using the closed form estimates of (θ, ϕ) stated above, the estimate of CACE is:

$$\begin{aligned}\hat{\delta}_{IV} &= \hat{\mu}_{T=1,C=1} - \hat{\mu}_{T=0,C=1}\\ &= \sum_{0}^{K-1} k \cdot \left(\frac{\hat{\theta}^{(1)}_{1k}}{\hat{\theta}^{(1)}_{1+}} - \frac{\hat{\theta}^{(0)}_{1k}}{\hat{\theta}^{(0)}_{1+}}\right)\\ &= \sum_{0}^{K-1} k \cdot \left(\frac{n_{(0),1k}}{n_{(0),1+}} - \frac{n_{(2),+k}(n_0+n_1) - n_{(0),0k}(n_2+n_3)}{n_2(n_0+n_1) - n_{(0),0+}(n_2+n_3)}\right) \qquad (3.15)\end{aligned}$$

which is the same as the MOM estimator in (3.11) proposed by Frangakis and Rubin (1999). Therefore, Frangakis and Rubin's MOM estimators are ML providing the estimate of nuisance parameter $\phi^{(0)}_1$ lies in the parameter space.

3.3.2 NCEC is assumed for Y

MOM estimator. Under the same missing-data mechanism as that specified in section 3.3.1, the estimator of the mean of outcomes for compliers in the active treatment group $\mu_{T=1,C=1}$ can still be estimated by (3.8). However, estimating $\mu_{T=0,C=1}$ is different from (3.10), as NCEC is assumed for Y. Under NCEC, the distributions of Y for compliers and never-takers do not differ in the control group, so that $P(Y = k|C = 1, T = 0) = P(Y = k|C = 0, T = 0) = P(Y = k|T = 0)$, $k = 0, \ldots, K-1$, and the mean of observed outcome in the control group can be used to estimate $\mu_{T=0}$:

$$\hat{\mu}_{T=0,C=1} = \hat{\mu}_{T=0,C=0} = \hat{\mu}_{T=0} = \hat{\mu}^{obs}_{T=0} \qquad (3.16)$$

As a result, we have the per-protocol estimator of the CACE which is restricted to participants who follow the protocol:

$$\begin{aligned}\hat{\delta}_{PP} &= \hat{\mu}_{T=1,C=1} - \hat{\mu}_{T=0} \\ &= \sum_{0}^{K-1} k \cdot (\frac{n_{(0),1k}}{n_{(0),1+}} - \frac{n_{(2),+k}}{n_2}) \end{aligned} \tag{3.17}$$

Non-iterative ML estimates. As in section 3.3.1, we can also find non-iterative ML estimates. After applying $\theta^{(0)}_{jk} = \theta^{(0)}_{j+} \cdot \theta^{(0)}_{+k}$ implied by Assumption 4, $\theta^{(0)}_{j+} = \theta^{(1)}_{j+}$ implied by Assumption 2 and $\phi^{(1)}_0 = \phi^{(0)}_0$ implied by Assumption 5, the closed-form estimators can be expressed as following:

$$\begin{aligned}\hat{\theta}^{(1)}_{jk} &= \frac{n_{(0),jk}}{n_{(0),j+}} \cdot \frac{n_{(0),j+} + n_{(1),j+}}{n_0 + n_1}, \\ \hat{\theta}^{(0)}_{jk} &= \frac{n_{(2),+k}}{n_2} \cdot \frac{n_{(0),j+} + n_{(1),j+}}{n_0 + n_1}, \\ \hat{\phi}^{(1)}_{j} &= \frac{n_{(1),j+}}{n_{(1),j+} + n_{(0),j+}}, \\ \hat{\phi}^{(0)}_{0} &= \hat{\phi}^{(1)}_{0} \\ \hat{\phi}^{(0)}_{1} &= \frac{\frac{n_3}{n_2+n_3} - \frac{n_{(1),0+}}{n_0+n_1}}{\frac{n_{(0),1+}+n_{(1),1+}}{n_0+n_1}}\end{aligned}$$

Except for $\hat{\theta}^{(0)}_{jk}$, these estimates are exactly the same as those obtained when ER is assumed for Y, and the corresponding estimate of the CACE is given by (3.17). Therefore, the MOM estimator of the CACE from the per-protocol analysis is the ML estimator providing $\hat{\phi}^{(0)}_1 \in [0,1]$. When $\hat{\phi}^{(0)}_1$ is not in $[0,1]$, the EM algorithm can be used to find ML estimates. The E-step is exactly the same in section 3.3.1, while M-step is different because the new assumptions affect the M-step.

3.4 Estimation when NCEC is assumed for M^Y

In this section, we assume latent ignorability along with a different assumption for M^Y. Under Assumption 6, missingness of Y in the treatment group could depend on the observed principal compliance C, whereas in the control group, where C is missing, the missingness of Y does not depend on C. In this case, the missing-data mechanism of Y is MAR.

3.4.1 ER is assumed for Y

MOM estimator. The missingness of Y in the active treatment group is not affected by the NCEC assumption for M^Y, so $\mu_{T=1,\,C=1}$ can still be estimated by (3.8) as in section 3.3. But since Y is also missing at random in the control group, $\mu_{T=0,\,C=1}$ has a different estimate. By Bayes' theorem and Assumptions 2, 3 and 6, we have

$$\mu_{T=0,\,C=1} = \frac{\mu_{T=0,\,M^Y=0} - \mu_{T=1,\,C=0}P(C=0)}{1-P(C=0)} \tag{3.18}$$

which subtracts never-taker outcomes from the overall observed outcomes in the control group. This expression makes use of the fact that never-taker outcomes are represented by outcomes of never-takers in the treated group. All quantities in the expression (3.18) can be directly estimated to obtain

$$\begin{aligned}\hat{\mu}_{T=0,\,C=1} &= \frac{\hat{\mu}_{T=0,\,M^Y=0} - \hat{\mu}_{T=1,\,C=0}\hat{P}(C=0)}{1-\hat{P}(C=0)} \\ &= \sum_{k=0}^{K-1} k \frac{\frac{n_{(2),+k}}{n_2} - \frac{n_{(0),0k}}{n_{(0),0+}}\frac{n_{(0),0+}+n_{(1),0+}}{n_0+n1}}{\frac{n_{(0),1+}+n_{(1),1+}}{n_0+n1}}\end{aligned} \tag{3.19}$$

By using estimates in (3.8) and (3.19), we then have for the CACE:

$$\begin{aligned}\hat{\delta}_{IV} &= \hat{\mu}_{T=1,\,C=1} - \hat{\mu}_{T=0,\,C=1} \\ &= \sum_{0}^{K-1} k\cdot(\frac{n_{(0),1k}}{n_{(0),1+}} - \frac{\frac{n_{(2),+k}}{n_2} - \frac{n_{(0),0k}}{n_{(0),0+}}\frac{n_{(0),0+}+n_{(1),0+}}{n_0+n_1}}{\frac{n_{(0),1+}+n_{(1),1+}}{n_0+n_1}})\end{aligned} \tag{3.20}$$

Non-iterative ML estimates. After applying $K+1$ constraints implied by Assumptions 2, 3 and 6, we have $\theta_{0k}^{(1)} = \theta_{0k}^{(0)}$, $k = 0, \ldots, K-1$ and $\phi_1^{(0)} = \phi_0^{(0)}$. Parameters in the model are just identified and the closed-form estimators are as following:

$$\begin{aligned}
\hat{\theta}^{(1)}_{jk} &= \frac{n_{(0),jk}}{n_{(0),j+}} \cdot \frac{n_{(0),j+} + n_{(1),j+}}{n_0 + n_1} \ , \\
\hat{\theta}^{(0)}_{0k} &= \hat{\theta}^{(1)}_{0k} \ , \\
\hat{\theta}^{(0)}_{1k} &= \frac{n_{(2),+k}}{n_2} - \frac{n_{(0),0k}}{n_{(0),0+}} \frac{n_{(0),0+} + n_{(1),0+}}{n_0 + n_1} \ , \\
\hat{\phi}^{(1)}_{j} &= \frac{n_{(1),j+}}{n_{(1),j+} + n_{(0),j+}} \ , \\
\hat{\phi}^{(0)}_{j} &= \frac{n_3}{n_2 + n_3}
\end{aligned}$$

Unlike closed-form estimates in section 3.3, the estimate of nuisance parameter $\phi^{(0)}_1$ is always in the parameter space $[0, 1]$, while the estimate of $\theta^{(0)}_{1k}$ may or may not lie in $[0, 1]$. Therefore, whether these closed-form estimates are ML estimates is determined by the values of $\hat{\theta}^{(0)}_{1k}$. If all of $\hat{\theta}^{(0)}_{1k}$ lie in $[0, 1]$, then they are ML estimates, and the estimate of CACE is:

$$\begin{aligned}
\hat{\delta}_{IV} &= \hat{\mu}_{T=1,C=1} - \hat{\mu}_{T=0,C=1} \\
&= \sum_{0}^{K-1} k \cdot (\frac{\hat{\theta}^{(1)}_{1k}}{\hat{\theta}^{(1)}_{1+}} - \frac{\hat{\theta}^{(0)}_{1k}}{\hat{\theta}^{(0)}_{1+}}) \\
&= \sum_{0}^{K-1} k \cdot (\frac{n_{(0),1k}}{n_{(0),1+}} - \frac{\frac{n_{(2),+k}}{n_2} - \frac{n_{(0),0k}}{n_{(0),0+}} \frac{n_{(0),0+}+n_{(1),0+}}{n_0+n_1}}{\frac{n_{(0),1+}+n_{(1),1+}}{n_0+n_1}}) \qquad (3.21)
\end{aligned}$$

which is the same as the MOM estimate in (3.20). Otherwise, the EM algorithm yields ML estimates. The E-step is exactly same as that in section 3.3, but the M-step is different.

3.4.2 NCEC is assumed for Y

MOM estimator. Similar to section 3.3.2, when NCEC is assumed for Y, $\mu_{T=0,C=1} = \mu_{T=0,C=0} = \mu_{T=0}$, and the mean of observed outcome in the control group can be used to estimate $\mu_{T=0}$:

$$\hat{\mu}_{T=0,C=1} = \hat{\mu}_{T=0,C=0} = \hat{\mu}_{T=0} = \hat{\mu}^{obs}_{T=0} \qquad (3.22)$$

As a result, we have the per-protocol estimator of the CACE:

$$\begin{aligned}\hat{\delta}_{PP} &= \hat{\bar{Y}}_{T=1,C=1} - \hat{\bar{Y}}_{T=0} \\ &= \sum_{0}^{K-1} k \cdot \left(\frac{n_{(0),1k}}{n_{(0),1+}} - \frac{n_{(2),+k}}{n_2}\right) \qquad (3.23)\end{aligned}$$

Non-iterative ML estimates. After applying $\theta^{(0)}_{jk} = \theta^{(0)}_{j+} \cdot \theta^{(0)}_{+k}$ implied by Assumption 4, $\theta^{(0)}_{j+} = \theta^{(1)}_{j+}$ implied by Assumption 2 and $\phi^{(0)}_0 = \phi^{(0)}_1$ implied by Assumption 6, our model is exactly identified and closed-form estimators can be calculated as follows:

$$\begin{aligned}\hat{\theta}^{(1)}_{jk} &= \frac{n_{(0),jk}}{n_{(0),j+}} \cdot \frac{n_{(0),j+} + n_{(1),j+}}{n_0 + n_1}, \\ \hat{\theta}^{(0)}_{jk} &= \frac{n_{(2),+k}}{n_2} \cdot \frac{n_{(0),j+} + n_{(1),j+}}{n_0 + n_1}, \\ \hat{\phi}^{(1)}_{j} &= \frac{n_{(1),j+}}{n_{(1),j+} + n_{(0),j+}}, \\ \hat{\phi}^{(0)}_{j} &= \frac{n_3}{n_2 + n_3}\end{aligned}$$

Under compound NCEC implied by Assumption 4 and 6, the estimates of both nuisance parameters and parameters of interest all lie in the parameter space, so non-iterative ML estimates always exist. As the consequence, the estimate of CACE can be computed as the following:

$$\begin{aligned}\hat{\delta}_{PP} &= \hat{\bar{Y}}_{T=1,C=1} - \hat{\bar{Y}}_{T=0} \\ &= \sum_{0}^{K-1} k \cdot \left(\frac{\hat{\theta}^{(1)}_{1k}}{\hat{\theta}^{(1)}_{1+}} - \hat{\theta}^{(0)}_{1k}\right) \\ &= \sum_{0}^{K-1} k \cdot \left(\frac{n_{(0),1k}}{n_{(0),1+}} - \frac{n_{(2),+k}}{n_2}\right) \qquad (3.24)\end{aligned}$$

which is the same as (3.23). So the MOM estimate is always the ML estimate when compound NCEC holds for Y and M^Y.

Summary of above analyses. Table 3.2 and 3.3 summarize all results under various assumptions. First, when NCEC is assumed for Y, no matter which assumption is applied for M^Y, NCEC or ER, we obtain the same non-iterative (or MOM) estimates of CACE, namely those which result from the PP analysis. However, the assumptions about M^Y determine whether or not the non-iterative (or MOM) estimate of CACE is the ML estimate. When NCEC is assumed for M^Y, the non-iterative (or MOM) estimate is always the ML estimate. When ER is assumed for M^Y, the non-iterative (or MOM) estimate is the ML estimate if $\hat{\phi}_1^{(0)} \in [0,1]$. On the other hand, when ER is assumed for Y, the different assumptions applied to M^Y result in different non-iterative (or MOM) IV estimates of CACE. When NCEC is assumed for M^Y, the non-iterative (or MOM) estimate is the ML estimate if all the estimates of parameters of interest $\hat{\theta}_{1k}^{(0)}$ fall in the interval $[0,1]$; while when ER is assumed for M^Y, the non-iterative (or MOM) estimate is the ML estimate if the estimate of nuisance parameter $\hat{\phi}_1^{(0)} \in [0,1]$.

3.5 Application

We calculate the PP and IV estimators for data from a double-blind clinical trial comparing clozapine versus haloperidol, two antipsychotic medications, in patients with refractory schizophrenia. Several clinical trials have shown that clozapine is more effective than other conventional antipsychotics, with fewer extrapyramidal side-effects (eg. stiffness, tremors, and other involuntary muscle movements). However, clozapine is more expensive and unfortunately associated with potentially fatal agranulocytosis which requires close monitoring and increases the cost. The current trial was conducted to compare the effectiveness and cost of clozapine with those of haloperidol, a widely used conventional treatment. The primary outcomes are symptoms of schizophrenia, quality of life, days in the hospital for psychiatric reasons, and costs.

We focus on analyzing the positive and negative syndrome score (PANSS), a measure of symptoms of schizophrenia. With possible scores from 30 to 210, higher values of PANSS indicate more severe symptoms. The trial has binary and continuous outcomes for PANSS. O'Malley and Normand (2005) calculated MOM and ML estimates by using a continuous PANSS score at 1-year follow-up, whereas Levy, O'Malley and Normand (2004) considered a covariate adjustment for a binary outcome, a 20% reduction in PANSS score which is considered as a clinically important improvement. Both of these papers computed the IV estimate of the ITT effect under assumptions of latent ignorability and compound ER. We consider MOM and ML estimates for PP and IV estimates of the CACE effect under varieties of assumptions, such as compound ER or compound NCEC as described above.

Table 3.4 summarizes the characteristics of the sample with the restricted access to clozapine. The 161 patients randomized to haloperidol did not have access to clozapine and had to take haloperidol. On the other hand, among 144 patients randomized to clozapine, there were 22 patients who switched to take haloperidol because of severe side-effects, lack of efficacy and non-drug-related reasons such as not wanting to continue the trial. Those who complied in the active treatment group had no missing data whereas, among the 22 non-compliers, there was a very high missingness rate of 60%.

Under various assumptions for both outcome and missingness of PANSS, estimates of parameters of interest, nuisance parameters and CACE are listed in Table 3.5. Standard errors are estimated by a bootstrap with 1000 bootstrap samples. If a bootstrap sample leads to 'estimates' that lie outside of the parameter space, the EM algorithm is used to obtain the ML estimates. With the 'estimates' of all parameters between 0 and 1, MOM estimates are ML estimates under these four different scenarios. It is interesting to note that the estimates of parameters $\theta_{jk}^{(1)}$ and $\phi_j^{(1)}$ in the clozapine group are exactly the same across four scenarios. Although these scenarios put different constraints on the parameters $\theta_{jk}^{(0)}$ and $\phi_j^{(0)}$ in the haloperidol group and therefore yield different estimates for them, $\theta_{jk}^{(1)}$ and $\phi_j^{(1)}$ can always be estimated by using the observed data in their corresponding group, because the outcome of PANSS will be completely missing at random given the observed compliance status in the clozapine group. When NCEC is assumed for PANSS, which means both compliers and never-takers in the haloperidol group have the same distribution of PANSS, assumptions of its missingness do not affect the estimate of CACE. With the information of never-takers in the clozapine group ignored, this estimate is the PP estimate. On the other hand, when ER is assumed for PANSS, which means never-takers in two groups have same distribution of PANSS, the information of all patients is used to estimate CACE, and these estimates are IV estimates. Moreover, IV estimates are influenced by the assumptions of the missingness of PANSS. Furthermore, for the estimates of CACE, the ignorable missing-data mechanism of PANSS (when NCEC is assumed) yields smaller standard errors than the non-ignorable missing-data mechanism of PANSS (when ER is assumed).

With strong dependence of estimates of CACE on assumptions of both outcome and missingness of PANSS, we advise caution and careful examination of these assumptions when analyzing and interpreting the data. Since compliers and never-takers have different rates of missing data and outcome distributions in the clozapine group, NCEC for PANSS and its missingness maybe not plausible here. However, it is still worth considering them, since they provide the foundation of PP and IV analysis.

3.6 Conclusions and future work

Analyzing randomized clinical trials involving human subjects is complicated by non-compliance and subsequent non-response, since, in addition to specifying the missing-data mechanism, we need to model both non-compliance and missing outcomes to identify the model. In this chapter, we discuss various assumptions for principal compliance and missing outcome in randomized clinical trials with a categorical outcome. We find the choice of PP and IV analysis depends on assumptions made about principal compliance. If there is no effect of the treatment assignment on the distribution of the outcome for never-takers, we can use IV analysis, while if the distributions of the outcome are assumed to be same for principal compliers and never-takers in the control group, PP analysis is used to estimate treatment efficacy. Furthermore, the reasons why the outcome has missing values should also be carefully evaluated, since they determine whether MOM estimates are ML estimates or not for each type of analysis.

We specify two assumptions for missing outcomes in randomized trials. Although it is impossible to test these assumptions, sensitivity analyses can be developed to evaluate their influences on the estimators of treatment efficacy. For example, for the ER assumption of missing outcome, sensitivity analysis can be carried out by defining a nuisance parameter as the ratio of proportions of missing outcome for never-takers between the active treatment group and the control group. The estimators of treatment efficacy are then functions of this nuisance parameter. By varying the nuisance parameter, we can assess the effect of violations of ER assumption for missing outcome on the estimators of treatment efficacy.

There are many possible generalizations and extensions of our methods:

(A) Models for data involving continuous or ordinal variables, with the same pattern as that described here.

(B) Models for clinical trials with partial compliance and/or two active treatment arms.

(C) Models for three-level principal compliance. Besides compliers and never-takers, there are also always-takers in randomized trials.

(D) Models for the additional of fully observed covariates to the data structure considered here. Compliance maybe associated with some covariates and estimates can then be obtained more precisely.

(E) Models for clinical trials with longitudinal setting considered by Peng, Little and Raghunathan (2004).

Table 3.2: Non-iterative Estimates of Parameters

Assumptions		**Parameter of Interest**			**Nuisance Parameter**		
Y	M^Y	$\theta_{jk}^{(1)}$	$\theta_{0k}^{(0)}$	$\theta_{1k}^{(0)}$	$\phi_j^{(1)}$	$\phi_0^{(0)}$	$\phi_1^{(0)}$
ER	ER	$\frac{n_{(0),jk}}{n_{(0),j+}} \cdot \frac{n_{(0),j+}+n_{(1),j+}}{n_0+n_1}$	$\hat{\theta}_{0k}^{(1)}$	$\frac{(n_0+n_1)n_{(2),+k}-(n_2+n_3)n_{(0),0k}}{(n_0+n_1)n_2-(n_2+n_3)n_{(0),0+}} \cdot \frac{n_{(0),1+}+n_{(1),1+}}{n_0+n_1}$	$\frac{n_{(1),j+}}{n_{(1),j+}+n_{(0),j+}}$	$\hat{\phi}_0^{(1)}$	$\frac{\frac{n_3}{n_2+n_3}-\frac{n_{(1),0+}}{n_0+n_1}}{\frac{n_{(0),1+}+n_{(1),1+}}{n_0+n_1}}$
ER	NCEC	*same*	$\hat{\theta}_{0k}^{(1)}$	$\frac{n_{(2),+k}}{n_2} - \frac{n_{(0),0k}}{n_{(0),0+}} \frac{n_{(0),0+}+n_{(1),0+}}{n_0+n_1}$	*same*	$\frac{n_3}{n_2+n_3}$	$\frac{n_3}{n_2+n_3}$
NCEC	NCEC	*same*	$\frac{n_{(2),+k}}{n_2} \cdot \frac{n_{(0),0+}+n_{(1),0+}}{n_0+n_1}$	$\frac{n_{(2),+k}}{n_2} \cdot \frac{n_{(0),1+}+n_{(1),1+}}{n_0+n_1}$	*same*	$\frac{n_3}{n_2+n_3}$	$\frac{n_3}{n_2+n_3}$
NCEC	ER	*same*	$\frac{n_{(2),+k}}{n_2} \cdot \frac{n_{(0),0+}+n_{(1),0+}}{n_0+n_1}$	$\frac{n_{(2),+k}}{n_2} \cdot \frac{n_{(0),1+}+n_{(1),1+}}{n_0+n_1}$	*same*	$\hat{\phi}_0^{(1)}$	$\frac{\frac{n_3}{n_2+n_3}-\frac{n_{(1),0+}}{n_0+n_1}}{\frac{n_{(0),1+}+n_{(1),1+}}{n_0+n_1}}$

Table 3.3: MLE of Treatment Efficacy

Assumptions		**Treatment Efficacy**
Y	M^Y	CACE
ER	ER	if $\hat{\phi}_1^{(0)} \in [0,1]$, MLE is MOM: $\hat{\delta}_{IV} = \sum_0^{K-1} k \cdot (\frac{n_{(0),1k}}{n_{(0),1+}} - \frac{n_{(2),+k}(n_0+n_1)-n_{(0),\,0k}(n_2+n_3)}{n_2(n_0+n_1)-n_{(0),\,0+}(n_2+n_3)})$ otherwise, using EM algorithm
ER	NCEC	if $\hat{\theta}_{1k}^{(0)} \in [0,1]$, MLE is MOM: $\hat{\delta}_{IV} = \sum_0^{K-1} k \cdot (\frac{n_{(0),1k}}{n_{(0),1+}} - \frac{\frac{n_{(2),+k}}{n_2} - \frac{n_{(0),0k}}{n_{(0),0+}} \frac{n_{(0),0+}+n_{(1),0+}}{n_0+n_1}}{\frac{n_{(0),1+}+n_{(1),1+}}{n_0+n_1}})$ otherwise, using EM algorithm
NCEC	NCEC	MLE is MOM: $\hat{\delta}_{PP} = \sum_0^{K-1} k \cdot (\frac{n_{(0),1k}}{n_{(0),1+}} - \frac{n_{(2),+k}}{n_2})$
NCEC	ER	$\hat{\phi}_1^{(0)} \in [0,1]$, MLE is MOM: $\hat{\delta}_{PP} = \sum_0^{K-1} k \cdot (\frac{n_{(0),1k}}{n_{(0),1+}} - \frac{n_{(2),+k}}{n_2})$ otherwise, using EM algorithm

Table 3.4: Randomized Clinical Trials with Clozapine vs. Haloperidol

	assigned clozapine took clozapine	assigned clozapine took haloperidol	assigned clozapine	assigned haloperidol
sample size:	122	22	144	161
missing rate:	0	0.60	0.10	0.30
fraction with 20% reduction in PANSS at 1 year	0.40	0.10	0.40	0.30

Table 3.5: Estimates of Parameters and Treatment Efficacy for Randomized Clinical Trials with Clozapine vs. Haloperidol

Assumptions		**Parameter of Interest**								**Nuisance Parameter**				**Treatment Efficacy**
Y	M^Y	$\theta_{00}^{(1)}$	$\theta_{01}^{(1)}$	$\theta_{10}^{(1)}$	$\theta_{11}^{(1)}$	$\theta_{00}^{(0)}$	$\theta_{01}^{(0)}$	$\theta_{10}^{(0)}$	$\theta_{11}^{(0)}$	$\phi_0^{(1)}$	$\phi_1^{(1)}$	$\phi_0^{(0)}$	$\phi_1^{(0)}$	CACE
ER	ER	0.134	0.019	0.493	0.354	0.134	0.019	0.571	0.276	0.636	0	0.636	0.237	0.092
		(0.025)	(0.019)	(0.040)	(0.039)	(0.025)	(0.019)	(0.044)	(0.041)	(0.080)	(0)	(0.080)	(0.005)	(0.067)
ER	NCEC	*same*				0.134	0.019	0.557	0.291	*same*		0.298	0.298	0.075
						(0.025)	(0.019)	(0.050)	(0.047)			(0.036)	(0.036)	(0.072)
NCEC	NCEC	*same*				0.105	0.047	0.585	0.262	*same*		0.298	0.298	0.108
						(0.015)	(0.009)	(0.039)	(0.037)			(0.036)	(0.036)	(0.063)
NCEC	ER	*same*				0.105	0.047	0.585	0.262	*same*		0.636	0.237	0.108
						(0.015)	(0.009)	(0.039)	(0.037)			(0.080)	(0.005)	(0.063)

3.7 Appendix

A1: Proof of equation (3.9)

Under latent ignorability, the expectation of observed outcome in the control group can be written as follows:

$$
\begin{aligned}
&\mu_{T=0,\,M^Y=0}\\
&= E[Y|T=0,\,M^Y=0]\\
&= E[Y|T=0,\,M^Y=0,\,C=1]P(C=1|T=0,\,M^Y=0)\\
&+ E[Y|T=0,\,M^Y=0,\,C=0]P(C=0|T=0,\,M^Y=0)\\
&= E[Y|T=0,\,C=1]P(C=1|T=0,\,M^Y=0)\\
&+ E[Y|T=0,\,C=0]P(C=0|T=0,\,M^Y=0)\\
&= \mu_{T=0,\,C=1}P(C=1|T=0,\,M^Y=0)+\mu_{T=0,\,C=0}P(C=0|T=0,\,M^Y=0)
\end{aligned}
$$

So the expectation of compliers' outcome in the control group can be expressed as:

$$
\mu_{T=0,\,C=1}=\frac{\mu_{T=0,\,M^Y=0}-\mu_{T=0,\,C=0}P(C=0|T=0,\,M^Y=0)}{P(C=1|T=0,\,M^Y=0)} \tag{3.25}
$$

While, for $P(C=j|T=0,\,M^Y=0)$, under assumption 2, we have:

$$
\begin{aligned}
P(C=j|T=0,\,M^Y=0)=&\frac{P(M^Y=0|T=0,\,C=j)P(C=j|T=0)}{P(M^Y=0|T=0)}\\
=&\frac{P(M^Y=0|T=0,\,C=j)P(C=j)}{P(M^Y=0|T=0)}
\end{aligned} \tag{3.26}
$$

After applying (3.26) into (3.25), we have:

$$
\begin{aligned}
&\mu_{T=0,\,C=1}\\
&=\frac{\mu_{T=0,\,M^Y=0}P(M^Y=0|T=0)-\mu_{T=0,\,C=0}P(M^Y=0|T=0,\,C=0)P(C=0)}{P(M^Y=0|T=0,\,C=1)P(C=1)}
\end{aligned} \tag{3.27}
$$

On the other hand, under assumption 2, we have:

$$
\begin{aligned}
&P(M^Y = 0|T = 0)\\
&= P(M^Y = 0|T = 0,\, C = 0)P(C = 0|T = 0)\\
&+ P(M^Y = 0|T = 0,\, C = 1)P(C = 1|T = 0)\\
&= P(M^Y = 0|T = 0,\, C = 0)P(C = 0) + P(M^Y = 0|T = 0,\, C = 1)P(C = 1)
\end{aligned}
$$

So, we have:

$$
\begin{aligned}
&P(M^Y = 0|T = 0,\, C = 1)P(C = 1)\\
&= P(M^Y = 0|T = 0) - P(M^Y = 0|T = 0,\, C = 0)P(C = 0)
\end{aligned} \tag{3.28}
$$

After plugging (3.28) into (3.27), we have:

$$
\begin{aligned}
&\mu_{T=0,\,C=1}\\
&= \frac{\mu_{T=0,\,M^Y=0}P(M^Y = 0|T = 0) - \mu_{T=0,\,C=0}P(M^Y = 0|T = 0,\, C = 0)P(C = 0)}{P(M^Y = 0|T = 0) - P(M^Y = 0|T = 0,\, C = 0)P(C = 0)}
\end{aligned} \tag{3.29}
$$

Under assumption 3 and 5, finally we have:

$$
\begin{aligned}
&\mu_{T=0,\,C=1}\\
&= \frac{\mu_{T=0,\,M^Y=0}P(M^Y = 0|T = 0) - \mu_{T=1,\,C=0}P(M^Y = 0|T = 1,\, C = 0)P(C = 0)}{P(M^Y = 0|T = 0) - P(M^Y = 0|T = 1,\, C = 0)P(C = 0)}
\end{aligned} \tag{3.30}
$$

A2: Proof of equation (3.18)

Proof is similar to A1, except in (3.26), under assumption 2 and 6, we have:

$$\begin{aligned}
P(C=j|T=0,\,M^Y=0)=&\frac{P(M^Y=0|T=0,\,C=j)P(C=j|T=0)}{P(M^Y=0|T=0)}\\
=&\frac{P(M^Y=0|T=0,\,C=j)P(C=j)}{\sum_{j=0}^{1}P(M^Y=0|T=0,C=j)P(C=j|T=0)}\\
=&\frac{P(M^Y=0|T=0)P(C=j)}{P(M^Y=0|T=0)\sum_{j=0}^{1}P(C=j|T=0)}\\
=&P(C=j)
\end{aligned} \tag{3.31}$$

After applying (3.31) into (3.25), we have:

$$\mu_{T=0,\,C=1}=\frac{\mu_{T=0,\,M^Y=0}-\mu_{T=0,\,C=0}P(C=0)}{P(C=1)}$$

Under assumption 3, finally we have:

$$\mu_{T=0,\,C=1}=\frac{\mu_{T=0,\,M^Y=0}-\mu_{T=1,\,C=0}P(C=0)}{1-P(C=0)}$$

CHAPTER IV

Combining Bootstrap and Bayes Inferences via Discrepancy Statistics

Abstract In the case of independent identically distributed samples, the simple bootstrap yields confidence limits that are asymptotically correct to the first order, but have less certain confidence coverage in small samples. Bayesian credibility intervals based on the posterior distribution of the model parameters tend to perform better for small samples, but are more dependent on modeling assumptions than the bootstrap. A discrepancy statistic based on the difference of model and bootstrap estimates of variance is used as a basis for combining bootstrap and Bayesian inferences. The goal is to achieve a compromise that combines the advantages of those two methods, yielding intervals that combine robustness with good small-sample confidence coverage. We assess properties of our method by some simple simulation experiments.

Keywords: Bayesian inference, robust inference, posterior predictive checks.

4.1 Introduction

Bootstrap methods (e.g. Efron, 1979, 1981, 1982) provide tools that can be used to set confidence intervals in complex problems. As the methods eliminate the routine but tedious theoretical calculations usually associated with precision assessment, they have increased the range of statistical problems that can be analyzed, and reduced the assumptions of the analysis. However, they perform poorly in some small sample problems, such as setting a confidence interval for the variance (Schenker, 1985). Bayesian credibility intervals based on the posterior distribution of the model parameters tend to perform better for small samples, but are more dependent on modeling assumptions than the bootstrap. In this article, a discrepancy statistic is introduced to combine bootstrap and Bayesian inferences, yielding intervals that are model robust with good small-sample confidence coverage.

4.2 Bootstrap confidence intervals

Let $\hat{\theta}$ be a consistent estimate of a scalar parameter θ based on a sample $S = \{y_i : i = 1, \ldots, n\}$ of independent observations. Let $S^{(b)}$ be a sample of size n obtained from the original sample S by simple random sampling with replacement, and let $\hat{\theta}^{(b)}$ be the estimate of θ obtained by applying the original estimation method to $S^{(b)}$, where b indexes the drawn samples. Let $\left(\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}\right)$ be the set of estimates obtained by repeating this procedure B times. The bootstrap estimate is

$$\hat{\theta}_{boot} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)}$$

Large-sample precision can be estimated from the bootstrap distribution of $\hat{\theta}^{(b)}, b = 1, \ldots, B$. In particular, the bootstrap estimate of the variance of $\hat{\theta}$ (or $\hat{\theta}_{boot}$) is

$$\hat{V}_{boot} = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\theta}^{(b)} - \hat{\theta}_{boot}\right)^2$$

Under certain conditions, it can be shown that (a) the bootstrap estimator $\hat{\theta}_{boot}$ is less biased than the original estimator $\hat{\theta}$, and under quite general conditions, (b) $\hat{V}_{boot}$ is a consistent estimate of the variance of $\hat{\theta}$ (or $\hat{\theta}_{boot}$) as n and B tend to infinity. If the bootstrap distribution is approximately normal, property (b) implies that an approximate $100(1-\alpha)\%$ bootstrap confidence interval for θ can be computed as

$$I_{norm}^{boot}(\theta) = \hat{\theta} \pm Z_{1-\alpha/2}\sqrt{\hat{V}_{boot}} \tag{4.1}$$

where $Z_{1-\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution. Alternatively, an approximate $100(1-\alpha)\%$ bootstrap confidence interval is given by

$$I_{emp}^{boot}(\theta) = \left(\hat{\theta}^{(b,\,l)},\, \hat{\theta}^{(b,\,u)}\right) \tag{4.2}$$

where $\hat{\theta}^{(b,\,l)}$ and $\hat{\theta}^{(b,\,u)}$ are the empirical $(\alpha/2)$ and $(1-\alpha/2)$ quantiles of the bootstrap distribution of $\hat{\theta}$. This interval may be preferable when the bootstrap distribution of $\hat{\theta}$ is not close to normal.

Stable intervals based on (4.1) require bootstrap samples of the order of $B = 200$. Intervals based on (4.2) require much larger bootstrap samples, for example $B = 2000$ or more (Efron, 1993).

Preliminary simulations for the situations we consider, concerning confidence intervals for the logarithm of population variance, suggested that confidence intervals based on (4.1) have better coverage probability than those based on (4.2), whether or not the distribution of population is normal. So in the remainder of this article we focus on bootstrap confidence intervals using (4.1).

4.3 Bayesian credibility intervals

The Bayesian analogue of a frequentist confidence interval (CI) is usually referred to as a credibility interval. Specifically, an interval I satisfying

$$P(\theta \in I|\mathbf{y}) = \int_I p(\theta|\mathbf{y})d\theta = 1 - \alpha$$

is called a $100(1-\alpha)\%$ credibility interval for θ. Integration is replaced by summation for discrete components of θ. Unlike the frequentist confidence interval, this definition provides direct probability statements about the probability that θ lies in I given the observed data $\mathbf{y}$, based on the observed likelihood and the prior $\pi(\theta)$.

In problems where θ has high dimension, the integrals involved in $p(\theta|\mathbf{y})$ may be very difficult to compute. This problem has been greatly reduced by stochastic simulation methods that take independent draws from $p(\theta|\mathbf{y})$. Let $\left(\theta^{(1)}, \ldots, \theta^{(D)}\right)$ represent the drawn values from $p(\theta|\mathbf{y})$ where D is very large. A $100(1-\alpha)\%$ credibility interval for θ is

$$I_{emp}^{bayes}(\theta) = \left(\theta^{(d,\,l)},\, \theta^{(d,\,u)}\right) \tag{4.3}$$

where $\theta^{(d,\,l)}$ and $\theta^{(d,\,u)}$ are $(\alpha/2)$ and $(1-\alpha/2)$ quantiles of the empirical distribution of the draws $\theta^{(d)}, d = 1, \ldots, D$. If the posterior distribution is close to normal, an approximate $100(1-\alpha)\%$ credibility interval for θ is

$$I_{norm}^{bayes}(\theta) = \tilde{\theta} \pm Z_{1-\alpha/2}\sqrt{\tilde{V}_{bayes}} \tag{4.4}$$

where $\tilde{\theta}$ and $\tilde{V}_{bayes}$ are the mean and variance of the simulated data $\theta^{(d)}, d = 1, \ldots, D$.

Suppose $y_1, \ldots, y_n$ is a random sample from a univariate normal distribution with mean μ and variance σ^2. With the conventional Jeffreys' prior distribution

$$p(\mu, \sigma^2) \propto 1/\sigma^2$$

the posterior distribution of μ and σ^2 is

$$\sigma^2|\mathbf{y} \sim nS^2/\chi^2_{n-1} \tag{4.5}$$

$$\mu|(\mathbf{y},\, \sigma^2) \sim N(\hat{\mu},\, \sigma^2/n) \tag{4.6}$$

where $\hat{\mu}$ and S^2 are maximum likelihood estimates of μ and σ^2 based on $y_1, \ldots, y_n$. Furthermore, the posterior distribution of $\theta = log(\sigma^2)$ is

$$\theta|\mathbf{y} \sim log(nS^2/\chi^2_{n-1})$$

From preliminary simulation results, we found that the Bayesian credibility intervals of $\theta = log(\sigma^2)$ based on (4.3) have coverage probabilities similar to those based on (4.4). For ease of comparison with the bootstrap confidence intervals from (4.1), we focus on Bayesian credibility intervals from (4.4) in the remainder of this chapter. Also, we assess the confidence coverage of the Bayesian credibility interval in repeated sampling, so our assessment is frequentist. Some Bayesians may question this tactic, but it does allow a direct comparison of the bootstrap and Bayesian intervals, and is consistent with the "calibrated Bayes" perspective on inference advocated by Box (1980), Rubin (1984) and others. For a recent discussion, see Little (2006).

4.4 Posterior predictive assessment of model fit via discrepancies

Assessing the plausibility of an assumed model is always important to avoid misleading inferences, so any meaningful inference should include a check that the assumed model is in agreement with the data. A classical approach calculates a tail-area probability under the assumed model to assess how extreme is the observed value of a goodness-of-fit test statistic. For some problems, such as linear models, goodness-of-fit tests are easy to implement since the reference distribution of the test statistic is known. Useful approximations to the null distribution of the test statistic can be found for other problems, but are not always available (see, for example, McCullagh 1985, 1986). In the Bayesian framework, posterior predictive model checking does not require known or approximately known reference distributions. Bayesian posterior predictive assessment was introduced in Guttman (1967), applied in Rubin (1981) and given a formal Bayesian definition in Rubin (1984). The idea is to measure departures of the observed data from the assumed model using the posterior predictive distribution of discrepancy measures $D(\mathbf{y})$. Gelman et al. (1996) considered more general measures $D(\mathbf{y}, \theta)$ that depend on parameters as well as data.

Let $\mathbf{y}^*$ be a future sample arising from the assumed model H given the observed data $\mathbf{y}$. With θ and $\mathbf{y}^*$ varying according to their joint posterior distribution, we compare $D(\mathbf{y}^*, \theta)$ with $D(\mathbf{y}, \theta)$ for the observed $\mathbf{y}$. The more extreme is the value of $D(\mathbf{y}, \theta)$, the greater is the evidence against the assumed model. A convenient summary measure of the discrepancy is the tail area probability

$$\begin{aligned} P_D &\equiv P\left[D(\mathbf{y}^*, \theta) \leq D(\mathbf{y}, \theta) | \mathbf{y}\right] \\ &= \int P\left[D(\mathbf{y}^*, \theta) \leq D(\mathbf{y}, \theta) | \theta\right] p(\theta|\mathbf{y})\, d\theta \\ &= \int\int I_{D(\mathbf{y}^*, \theta) \leq D(\mathbf{y}, \theta)} p(\mathbf{y}^*|\theta) p(\theta|\mathbf{y})\, d\mathbf{y}^* d\theta \end{aligned}$$

which is the classical p-value averaged over the posterior distribution of θ. This is the p-value defined by Rubin (1984), which we term the posterior predictive p-value (also see Meng, 1994 and Gelman et al, 1996) to contrast it with the prior predictive p-value of Box (1980). Some authors (Robins, van der Vaart and Ventura, 2000; Bayarri and Berger, 2000) have criticized the posterior predictive p-value since it does not in general have a uniform distribution under H, but proponents argue that it remains a valid measure from a Bayesian perspective; we make pragmatic use of the measure here, without attempting to resolve that controversy.

It is straightforward to estimate the posterior predictive p-value by the following simulations:

1. Draw θ^j from the posterior distribution of θ given $\mathbf{y}$;
2. For given θ^j, draw a predicted value $\mathbf{y}^{*,j}$ from the sampling distribution $P(\mathbf{y}^*|\theta^j)$;
3. Calculate $D(\mathbf{y}^{*,j}, \theta^j)$ and $D(\mathbf{y}, \theta^j)$;

A scatter plot of $\{(D(\mathbf{y}^{*,j}, \theta^j),\ D(\mathbf{y}, \theta^j)), j = 1, \ldots, J\}$ provides graphical assessments, and P_D is estimated by the proportion of the J pairs for which $D(\mathbf{y}, \theta^j) \geq D(\mathbf{y}^{*,j}, \theta^j)$.

4.5 Combining Bootstrap and Bayesian intervals

The confidence coverage of the bootstrap confidence interval for $\theta = log(\sigma^2)$ based on (4.1) can be compared with confidence coverage of the Bayesian credibility interval based on (4.4), for the normal model $N(\mu, \sigma^2)$ with a Jeffreys' prior for (μ, σ^2). Theory suggests, and simulations confirm, that under a correctly specified model, the Bayesian credibility intervals are similar to bootstrap confidence intervals in large samples, and are superior for small samples. On the other hand when the assumed model is far from the true model, Bayesian credibility intervals are inferior, particularly in large samples when bias from model misspecification dominates. To achieve a compromise that combines the best features of these two intervals, we define a discrepancy statistic $D(\mathbf{y})$ which is the ratio of model and bootstrap estimates of variance

$$D(\mathbf{y}) \equiv \frac{\tilde{V}_{bayes}(\theta)}{\hat{V}_{boot}(\hat{\theta})}$$

where $\tilde{V}_{bayes}(\theta)$ and $\hat{V}_{boot}(\hat{\theta})$ are Bayesian and bootstrap estimates of variance. If the model is correctly specified, this ratio is around 1, in large samples; otherwise, it may be smaller or larger than 1, depending on the form of model misspecification. By comparing the observed value of the discrepancy $D(\mathbf{y})$ with its posterior predictive distribution, we can assess whether the posited model fits the observed data set. Given the observed data $\mathbf{y}$ and the posited model which is normal with Jeffreys' prior, the posterior predictive distribution of the discrepancy is constructed as described before. For $j = 1, \ldots, J$,

1. Draw $\sigma^{2,j}$ and μ^j from (4.5) and (4.6);
2. Given μ^j and $\sigma^{2,j}$, draw a predictive data set $\mathbf{y}^{*,j}$ from $N(\mu^j, \sigma^{2,j})$;
3. For given $\mathbf{y}^{*,j}$, find $\tilde{V}^{*,j}_{bayes}(\theta)$ and $\hat{V}^{*,j}_{boot}(\hat{\theta})$, then calculate $D(\mathbf{y}^{*,j})$;

Note $\tilde{V}^{*,j}_{bayes}(\theta)$ is the estimated posterior variance of θ given the data set $\mathbf{y}^{*,j}$; and $\hat{V}^{*,j}_{boot}(\hat{\theta})$ is the bootstrap estimate of variance for $\hat{\theta}$ based on the data set $\mathbf{y}^{*,j}$.

Under the posterior predictive distribution of the discrepancy, we can calculate a posterior predictive p-value to quantify how extreme is the observed value of the discrepancy. If the observed value of the discrepancy $D(\mathbf{y})$ does not fall in the tail of the posterior predictive distribution, the posited model is deemed to fit the observed data, and the Bayesian credibility interval is used for inference; if the observed value of the discrepancy $D(\mathbf{y})$ does fall in the tail of the posterior predictive distribution, the posited model is deemed not to fit the observed data, and a confidence interval is constructed by combining the bootstrap confidence interval and the Bayesian credibility interval via a function of posterior predictive p-value P_D. Specifically, we define a weighted bootstrap/Bayes (WBB) $100(1-\alpha)\%$ confidence interval for the parameter of interest as follows

$$W(P_D) * I^{bayes}_{norm} + [1 - W(P_D)] * I^{boot}_{norm}$$

$$where : W(P_D) = \begin{cases} P_D & \text{if } P_D < 0.05 \\ 1 & \text{if } P_D \geq 0.05 \end{cases}$$

$$P_D = P\left(D(\mathbf{y}^*) \leq D(\mathbf{y}) | \mathbf{y}\right)$$

We apply this approach to the problem of estimating the logarithm of variance $\theta = log(\sigma^2)$ for data $\mathbf{y} = \{y_1, \ldots, y_n\}$ sampled from a t distribution, when the assumed model is normal with a Jeffreys' prior. The assumed normal model is deemed not to fit the data when the observed discrepancy falls in the left tail of its posterior predictive distribution. To compare the performance of the WBB method with the bootstrap confidence and Bayesian credibility intervals, a simulation study was conducted for samples of size $15, 20, 25, 30$ and 50. For each sample size, 10,000 data sets were simulated from a t distribution with degree of freedom 4 and true $\theta = log(\sigma^2) = 0.6931$. In computing the bootstrap confidence interval, $B = 200$ bootstrap replications were used. For the stochastic simulation of posterior distributions, $D = 10,000$ draws were used and the posterior predictive distribution of $D(\mathbf{y}^*)$ was simulated using $J = 10,000$. For every simulated data set $\mathbf{y}$, the left tail-area probability P_D was calculated to quantify the extremeness of the observed value of the discrepancy. The $10,000$ simulated data sets were stratified into three groups based on whether $P(D) < 0.05$, $0.05 \leq P(D) \leq 0.95$ and $P(D) > 0.95$. The average intervals and proportions of the intervals covering $\theta = log(\sigma^2) = 0.6931$ are given in Table 4.1 for each stratum and overall.

Comparing the bootstrap and Bayes intervals, we see that the overall coverage of the bootstrap intervals is below nominal, particular in small samples; this is consistent with the results in Schenker (1985). The Bayes intervals also have poor confidence coverage, particular for the larger sample sizes. An interesting feature is that the bootstrap confidence coverages vary much more than those of the Bayes intervals across the strata defined by the discrepancy statistic. This is related to the fact that Bayes is more “conditional” and hence less sensitive to this ancillary statistic. The intervals based on WBB have the best overall coverage rates of the three methods, suggesting that WBB combines the advantages of bootstrap and Bayesian

Table 4.1: 95% confidence interval and coverage probability for one tailed version of WBB when model is misspecified

sample	**method**	**left**	**middle**	**right**	**overall**
size		**tail**	**area**	**tail**	
15		(n=2403)	(n=7389)	(n=208)	(n=10000)
	bootstrap	(-0.50, 2.34)	(-0.42, 1.14)	(-0.36, 0.57)	(-0.44, 1.41)
		97.54%	77.20%	40.38%	81.32%
	bayes	(0.22, 1.76)	(-0.34, 1.20)	(-0.59, 0.95)	(-0.21, 1.33)
		73.24%	81.77%	72.12%	79.52%
	WBB	(-0.50, 2.33)	(-0.34, 1.20)	(-0.59, 0.95)	(-0.38, 1.47)
		97.34%	81.77%	72.12%	**85.31**%
20		(n=2961)	(n=6883)	(n=156)	(n=10000)
	bootstrap	(-0.30, 2.06)	(-0.28, 1.06)	(-0.26, 0.56)	(-0.28, 1.35)
		96.72%	77.57%	35.90%	82.59%
	bayes	(0.28, 1.59)	(-0.21, 1.10)	(-0.45, 0.85)	(-0.07, 1.24)
		74.64%	80.82%	70.51%	78.83%
	WBB	(-0.30, 2.06)	(-0.21, 1.10)	(-0.45, 0.85)	(-0.24, 1.38)
		96.72%	80.82%	70.51%	**85.37**%
25		(n=3473)	(n=6431)	(n=96)	(n=10000)
	bootstrap	(-0.19, 1.87)	(-0.20, 0.99)	(-0.13, 0.62)	(-0.19, 1.29)
		96.75%	77.71%	42.71%	82.70%
	bayes	(0.30, 1.46)	(-0.14, 1.02)	(-0.29, 0.87)	(-0.01, 1.17)
		74.52%	78.46%	69.79%	77.01%
	WBB	(-0.19, 1.87)	(-0.14, 1.02)	(-0.29, 0.87)	(-0.16, 1.31)
		96.66%	78.46%	69.79%	**84.70**%
30		(n=4039)	(n=5872)	(n=89)	(n=10000)
	bootstrap	(-0.11, 1.76)	(-0.13, 0.96)	(-0.10, 0.60)	(-0.13, 1.28)
		96.66%	74.63%	39.33%	83.21%
	bayes	(0.33, 1.38)	(-0.08, 0.97)	(-0.24, 0.81)	(0.09, 1.13)
		74.80%	76.26%	66.29%	75.58%
	WBB	(-0.11, 1.76)	(-0.08, 0.97)	(-0.24, 0.81)	(-0.09, 1.29)
		96.56%	76.26%	66.29%	**84.37**%
50		(n=5581)	(n=4394)	(n=25)	(n=10000)
	bootstrap	(0.06, 1.45)	(0.01, 0.86)	(-0.05, 0.53)	(0.03, 1.19)
		95.72%	71.48%	28.00%	84.90%
	bayes	(0.37, 1.17)	(0.05, 0.85)	(-0.14, 0.66)	(0.23, 1.03)
		75.08%	71.78%	44.00%	73.55%
	WBB	(0.06, 1.45)	(0.05, 0.85)	(-0.14, 0.66)	(0.05, 1.18)
		96.66%	71.78%	44.00%	**85.04**%

techniques. However, in this problem WBB does not achieve the nominal level of 95%. This is not surprising, given the deficiencies of two methods being combined: the bootstrap is asymptotically valid only to the first order, and the Bayesian interval is vulnerable to model misspecification. For those data with the observed discrepancy falling in the left tail of the posterior predictive distribution ($P_D < 0.05$), the bootstrap intervals have better coverage rates than Bayesian credibility intervals since the assumed model does not fit the data; while for those with the observed discrepancy in the middle area and right tail ($P_D \geq 0.05$), the Bayesian credibility intervals perform better. The choice of the left tail of the posterior predictive distribution of the discrepancy to assess whether the posited model matches the data assumes that the analyst knows that potential departures from the normal distribution are in the form of longer-tailed distributions. If longer or shorter than normal tails are entertained, the WBB method could be defined to measure discrepancies in both tails, as follows:

An approximate $100(1-\alpha)\%$ confidence interval of $\theta = log(\sigma^2)$ is

$$W(P_D) * I_{norm}^{bayes} + [1 - W(P_D)] * I_{norm}^{boot}$$

$$where : W(P_D) = \begin{cases} P_D & \text{if } P_D < \beta \\ 1 & \text{if } \beta \leq P_D \leq 1-\beta \\ 1 - P_D & \text{if } P_D > 1-\beta \end{cases}$$

$$P_D = P\left(D(\mathbf{y}^*) \leq D(\mathbf{y}) | \mathbf{y}\right)$$

Where β denotes the cut-off level to assess the extremeness of the observed value of the discrepancy under its posterior predictive distribution.

To check the validity of this WBB method and decide the value of β, additional simulations were done (Table 4.2). The value of β was taken as $0.025, 0.05, 0.10$ and 0.15. As before, the intervals based on WBB have the best coverage rates.

The coverage probabilities of WBB method do not depend much on the value of β. Although the choice of $\beta = 0.10$ or $\beta = 0.15$ appears slightly better for the cases considered, in subsequent simulations, we choose $\beta = 0.10$.

Methods were also compared for repeated samples of size 15 to 50 simulated from the normal distribution $N(0, 2)$ with $\theta = log(\sigma^2) = 0.6931$, to assess the performance of WBB when the model is correctly specified. As one might expect, the Bayes credibility intervals have the best confidence coverage in this situation, but WBB is only slightly worse, and does substantially improve the performance of the bootstrap confidence intervals in small samples.

Table 4.2: 95% confidence interval and coverage probability for two tailed version of WBB when model is misspecified

	method					
sample size	**bootstrap**	**bayes**	**WBB** (β=0.025)	**WBB** (β=0.05)	**WBB** (β=0.10)	**WBB** (β=0.15)
15	(-0.44, 1.39) 81.19%	(-0.22, 1.32) 81.03%	(-0.35, 1.43) 85.00%	(-0.37, 1.44) 85.36%	(-0.40,1.45) 85.45%	(-0.40, 1.45) 85.19%
20	(-0.29, 1.34) 81.79%	(-0.07, 1.23) 77.66%	(-0.23, 1.36) 83.37%	(-0.24, 1.37) 84.02%	(-0.26,1.38) 84.37%	(-0.27, 1.38) 84.32%
25	(-0.19, 1.30) 82.20%	(0.02, 1.17) 76.13%	(-0.13, 1.30) 83.12%	(-0.15, 1.31) 83.67%	(-0.17,1.32) 84.15%	(-0.18, 1.32) 84.26%
30	(-0.12, 1.27) 83.26%	(0.09, 1.13) 75.65%	(-0.08, 1.27) 83.64%	(-0.09, 1.28) 84.14%	(-0.10,1.29) 84.60%	(-0.11, 1.29) 84.75%
50	(0.03, 1.19) 84.49%	(0.23, 1.03) 73.26%	(0.07, 1.18) 83.60%	(0.05, 1.19) 84.45%	(0.04, 1.19) 84.98%	(0.04, 1.19) 85.15%

Table 4.3: 95% confidence interval and coverage probability for two tailed version of WBB when model is specified correctly

	method		
sample size	**bootstrap**	**bayes**	**WBB** (β=0.10)
15	(0.17, 1.63) 88.28%	(-0.07, 1.46) 95.02%	(-0.03, 1.52) **93.51**%
20	(0.22, 1.46) 89.41%	(0.04, 1.35) 95.18%	(0.07, 1.39) **93.82%**
25	(0.26, 1.35) 90.36%	(0.11, 1.27) 94.74%	(0.14, 1.29) **93.65%**
30	(0.29, 1.29) 91.12%	(0.17, 1.22) 95.54%	(0.19, 1.24) **94.60%**
50	(0.36, 1.14) 91.96%	(0.30, 1.10) 94.97%	(0.31, 1.11) **94.12%**

4.6 Discussion

We have proposed WBB as a method to combine bootstrap and Bayesian inferences through a function of the discrepancy and its posterior predictive p-value. For the problem of inference for the logarithm of variance, WBB generates intervals with better confidence coverage than the bootstrap when the model is correctly specified, and better confidence coverage than both bootstrap and Bayes when the model is misspecified. Besides improved performance for small samples, WBB is also asymptotically correct to the first order. Asymptotically under a misspecified model, the observed discrepancy has a posterior predictive p-value that tends to zero, and the WBB interval converges to the bootstrap interval.

In this chapter, we applied WBB to construct confidence intervals for the logarithm of population variance, with discrepancy defined as the ratio of Bayesian estimate of variance divided by bootstrap estimate of variance. The WBB can be applied to build confidence intervals for any parameter, given a suitable discrepancy (which might depend on parameters) to measure the differences of bootstrap and Bayesian inferences. The further development and assessment of the performance of WBB in other problems remains a topic for future research.

The weight function we considered here is one of many plausible choices, and the cut-off point for the posterior predictive p-value was set based on simulation results. More work is needed to evaluate other choices of weight functions and other choices of cut-offs for the posterior predictive p-value.

The WBB method is a plausible compromise between two simple Bayesian modeling and bootstrap strategies, and Bayesians and frequentists might both argue that better approaches are available. In particular, the naive bootstrap might be replaced by a more advanced bootstrap method, such as the BC_α or ABC method, or the studentized estimating function bootstrap. From a Bayesian perspective, a more principled approach when faced with an extreme value of the discrepancy statistic would be to modify the model to improve the fit – in the problem discussed here, replacing the normal model by the t model would obviously be the right approach, although of course in real settings we do not know the family from which the data are sampled. Such an approach requires more sophisticated modeling than our WBB method, which can be implemented in a relatively automated manner. Comparisons with these more sophisticated methods would also be of interest.

CHAPTER V

Conclusions and Future Work

The missing-data mechanism can be ignored when the data are MAR, but this assumption is not always intuitive for general pattern missing data. We consider a NMAR model, called AMAR, that is close to MAR and realistic in some settings. In randomized clinical trials subject to noncompliance to the treatment assignments and subsequent non-responses, we find AMAR is connected with latent ignorability proposed by Frangakis and Rubin (1999).

We examine the AMAR model in chapter II, for the case of bivariate categorical data and show that non-iterative ML estimates exist when 'estimates' of nuisance parameters fall in the parameter space. Although some data are discarded, estimates of the parameters of interest are still consistent and fully efficient. We also discuss extensions of this type of mechanism and develop likelihood ratio tests for AMAR and its nested models.

For randomized clinical trials with non-compliance and subsequent non-response considered in chapter III, under AMAR (or latent ignorability), we discuss various assumptions for the principal compliance and missing outcome. In each scenario defined by combinations of these assumptions, we derive ML estimates by using the EM algorithm, as well as non-iterative ML estimates by implementing pattern-mixture models with covariates, and find MOM estimates sometimes are ML estimates. We find assumptions of principal compliance decide which analysis is used to estimate the treatment efficacy. When ER satisfies, we use the IV analysis, while when NCEC satisfies, PP analysis is used to estimate treatment efficacy. Assumptions of missing outcome further determine whether MOM estimates are ML estimates or not.

Through a function of the discrepancy and its posterior predictive p-value, we propose WBB as a method to combine bootstrap and Bayesian inferences in chapter IV, to yield robust confidence intervals with good small-sample confidence coverages. For inference for the logarithm of the variance, we show that, no matter whether the model is correctly specified or not, WBB always generates intervals with better confidence coverage than the bootstrap, and when the model is misspecified, WBB generates better confidence coverage than Bayes. Besides improved performance for small samples, WBB is also asymptotically correct to the first order. Asymptotically under a misspecified model, the observed discrepancy has a posterior predictive p-value that tends to zero, and the WBB interval converges to the bootstrap interval.

We consider the AMAR model for bivariate categorical data in chapter II and III. In the future, it is worth extending models for bivariate data involving continuous or ordinal variables, with the same pattern and mechanism as that described here. It is also worthwhile to consider an extension that the first variable is binary and the second variable is normal with different means depending on the first binary variable. It remains an open question how to extend this type of model for data with more than two variables.

We consider one extension of the AMAR model with a fully observed covariate. Specifically, in the randomized clinical trials with non-compliance and non-response, the treatment assignment is a fully observed covariate, besides missing non-compliance and missing outcomes. If the AMAR model involves more than one fully observed covariates, the definition of AMAR should be carefully specified to incorporate the information of these observed covariates.

We study the randomized clinical trials with two-level compliance, compliers or never-takers, an extension to include always-takers is straightforward by modifying the categorical distribution of compliance in the model.

The compliance we consider here is all-or-none compliance. We will carefully extend our consideration to partial compliance as more restrictions are needed to identify the parameters in the model.

We specify various assumptions for compliance and missing outcomes in randomized trials. Although it is impossible to test these assumptions, sensitivity analyses can be developed to evaluate their influences on the estimators of treatment efficacy. For example, in the sensitivity analysis for the ER assumption of missing outcome, we can define a nuisance parameter as the ratio of proportions of missing outcomes for never-takers between the active treatment group and the control group. By varying this nuisance parameter, we can then assess the influences of the ER assumption of missing outcome on the estimators of treatment efficacy.

We consider randomized trials with two treatments, active treatment verses control treatment. Estimating treatment efficacy in randomized trials with more than two treatments (such as two active treatments and one control treatment) is more complicated, since it consists of more principal compliance categories and involves more complicated identifiability assumptions. We will study this extension in the future.

We build adaptive confidence intervals for the logarithm of population variance. Actually, WBB can be applied to build confidence intervals for any parameter, given a suitable discrepancy (which might depend on parameters) to measure the differences of bootstrap and Bayesian inferences. The further development and assessment of the performance of WBB in other problems remains a topic for future research. The weight function considered here is one of many plausible choices, and the cut-off point for the posterior predictive p-value is set based on simulation results. More work is needed to evaluate other choices of weight functions and cut-offs for the posterior predictive p-value. We constructed WBB method based on two simple Bayesian modeling and bootstrap strategies. In the future, it is possible to replace the naive bootstrap by a more advanced bootstrap method, such as the BC_α or ABC method, or the studentized estimating function bootstrap, and modify the model to improve the fit from a Bayesian perspective.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of Causal Effects Using Instrumental Variables (with Discussion). *J. Am. Statist. Assoc.* 91, 444-472.

[2] Baker, S. G. (1995) Marginal Regression for Repeated Binary Data with Outcome Subject to Non-ignorable Nonresponse. *Biometrics.* 51, 1042-1052.

[3] Baker, S. G. (1997) Compliance, All-or-None. In Kotz S, Read CR, Banks DL (eds.) *The Encyclopedia of Statistical Science.* Updated Volumn 1, 134-138. New York: John Wiley and Sons.

[4] Baker, S. and Laird, N. (1985) Categorical Response Subject to Nonresponse. *Technical Report, Department of Biostatistics, Harvard School of Public Health, Boston, MA.*

[5] Baker, S. G. and Lindeman, K. S. (1994) The Paired Availability Design: A Proposal for Evaluating Epidural Analgesia During Labor. *Statis. Med.* 13, 2269-2278.

[6] Barnard, J. and Rubin, D.B. (1999) Small-sample Degrees of Freedom with Multiple Imputation. *Biometrika* 86, 949-955.

[7] Bayarri, M.J. and Berger, J.O. (2000) P Values for Composite Null Models (with discussion). *J. Am. Statist. Assoc.* 95, 1127-1172.

[8] Birmingham, J. and Fitzmaurice, G. M. (2002) A Pattern-Mixture Model for Longitudinal Binary Responses with Nonignorable Nonresponse. *Biometrics.* 58, 989-996.

[9] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) Discrete Multivariate Analysis: Theory and Practice. *Cambridge, MA: MIT Press.*

[10] Box, G.E.P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *J.Roy. Statist. Soc. Ser. A.* 143, 383-430.

[11] Carlin, B.P. and Louis, T.A. (2000) Bayes and Empirical Bayes Methods for Data Analysis, second edition.

[12] Chen, T. and Fienberg, S. E. (1974) Two-dimensional Contingency Tables with Both Completely and Partially Classified Data. *Biometrics.* 30, 629-642.

[13] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum Likelihood from Incomplete Data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. B.* 39, 1-38.

[14] Durbin, J. (1954) Errors in Variables. *Rev. Int. Statist. Inst.* 22, 23-32.

[15] Ekholm, A. and Skinner, C. (1998) The Muscatine Children's Obesity Data Reanalysed Using Pattern Mixture Models. *Appl. Statist..* 47, 251-263.

[16] Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.* 7, 1-26.

[17] Efron, B. (1981) Nonparametric Standard Errors and Confidence Intervals. *Canadian Journal of Statistics* 9, 139-172.

[18] Efron, B. (1982) The Jackknife, the Bootstrap, and Other Resampling Plans. *National Science Foundation-Conference Board of the Mathematical Sciences Monograph* 38, Philadelphia: Society for Industrial and Applied Mathematics.

[19] Efron, B. and Tibshirani, R. (1993) An Introduction to the Bootstrap. New York:CRC Press.

[20] Fay, R. E. (1986) Causal Models for Patterns of Nonresponse. *J. Am. Statist. Assoc.* 81, 354-365.

[21] Frangakis, C. E. and Rubin, D. B. (1999) Addressing Complications of Intention-to-Treat Analysis in the Combined Presence of All-or-None Treatment-Noncompliance and Subsequent Missing Outcomes. *Biometrika.* 86, 365-379.

[22] Frangakis, C. E. and Rubin, D. B. (2002) Principal Stratification in Causal Inference. *Biometrics.* 58, 21-29.

[23] Fuchs, C. (1982) Maximum Likelihood Estimation and Model Selection in Contingency Tables with Missing Data. *J. Am. Statist. Assoc.* 77, 270-278.

[24] Gelman, A., Carlin, J.B., Stern, H. and Rubin, D.B. (2004) Bayesian Data Analysis. second edition.

[25] Gelman, A., Meng, X.L. and Stern, H. (1996) Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica* 6, 733-807.

[26] Glynn, R.J, Laird, N.M. and Rubin, D.B. (1986) Selection Modeling versus Mixture Modeling with Nonignorable Nonresponse. Drawing Inferencse From Self-Selected Samples (H. Wainer, ed.), 115-142.

[27] Glynn, R.J, Laird, N.M. and Rubin, D.B. (1993) Multiple Imputation in Mixture Models for Nonignorable Nonresponse with Follow-ups. *J. Am. Statist. Assoc.* 88, 984-993.

[28] Goldberger, A. S. (1972) Structral Equation Methods in the Social Sciences. *Econometrica.* 40, 979-1001.

[29] Goodman, L. A. (1970) The Multivariate Analysis of Qualitative Data: Interaction Among Multiple Classifications. *J. Am. Statist. Assoc.* 65, 225-256.

[30] Guttman, I.(1967) The Use of the Concept of a Future Observation in Goodness-Of-Fit Problems. *J.Roy. Statist. Soc. Ser. B* 29, 83-100.

[31] Hartley, H. O. (1958) Maximum Likelihood Estimation from Incomplete Data. *Biometrics.* 14, 174-194.

[32] Ibrahim, J.G. (1990) Incomplete Data in Generalized Linear Models. *J. Am. Statist. Assoc.* 85, 765-769.

[33] Ibrahim, J.G., Chen, M.H. and Lipsitz, S.R. (1999) Monte Carlo EM for missing covariates in parametric regression models. *Biometrics.* 55, 591-596.

[34] Levy, D. E., O'Malley, J. and Normand, SL. T. (2004) Covariate Adjustment in Clinical Trials with Non-ignorable Missing Data and Non-compliance. *Statis. Med.*. 23, 2319-2339.

[35] Lipsitz, S.R. and Ibrahim, J.G. (1996) A conditional model for incomplete covariates in parametric regression models. *Biometrika.* 83, 916-9222.

[36] Lipsitz, S.R., Ibrahim, J.G. and Fitzmaurice, G.M. (1999) Likelihood Methods for Incomplete Longitudinal Binary Responses with Incomplete Categorical Covariates. *Biometrics.* 55, 214-223.

[37] Lipsitz, S. R., Ibrahim, J. G. and Zhao, L. P. (1999) A Weighted Estimating Equation for Missing Covariate Data with Properites Similar to Maximum Likelihood. *J. Am. Statist. Assoc.* 94, 1147-1160.

[38] Lipsitz, S. R., Parzen, M. and Molenberghs, G. (1998) Obtaining the Maximum Likelihood Estimates in Incomplete R×C Contingency Tables Using a Poisson Generalized Linear Model. *J. Comp. and Graph. Statist.*. 7, 356-376.

[39] Little, R. J. A. (1982) Models for Nonresponse in Sample Surveys. *J. Am. Statist. Assoc.* 77, 237-250.

[40] Little, R. J. A. (1985b) Nonresponse Adjustments in Longitudinal Surveys: Models for Categorical Data. *Bulletin Int. Statist. Inst.* 15(1), 1-15.

[41] Little, R. J. A. (1993) Pattern-Mixture Models for Multivariate Incomplete Data. *J. Am. Statist. Assoc.* 88, 125-134.

[42] Little, R. J. A. (1994) A Class of Pattern-Mixture Models for Normal Incomplete Data. *Biometrika.* 81, 471-483.

[43] Little, R. J. A. (1995) Modeling the Drop-Out Mechanism in Repeated-Measures Studies. *J. Am. Statist. Assoc.* 90, 1112-1121.

[44] Little, R.J. (2006), Calibrated Bayes: A Bayes/Frequentist Roadmap. *The American Statistician.* 60, 213-223.

[45] Little, R. J. A., Long, Q. and Lin, X.H. (2008) A Comparison of Methods for Estimating the Causal Effect of a Treatment in Randomized Clinical Trials Subject to Noncompliance. *Biometrics.* In press.

[46] Little, R. J. A. and Rubin, D. B. (2002) Statistical Analysis with Missing Data.

[47] Little, R. J. A. and Wang, Y. X. (1996) Pattern-Mixture Models for Multivariate Incomplete Data with Covariate. *Biometrics.* 52, 98-111.

[48] Liu, C.H. and Rubin, D.B. (1994) The ECME Algorithm: A Simple Extension of EM and ECM with Fast Monotone Convergence. *Biometrika.* 81, 633-648.

[49] McCullagh, P. (1985) On the Asymptotic Distribution of Pearson's Statistics in Linear Exponential Family Models. *Internat. Statist. Rev.* 53, 61-67.

[50] McCullagh, P. (1986) The Condition Distribution of Goodness-of-Fit Statistics for Discrete Data. *J. Amer. Statist. Assoc.* 81, 104-107.

[51] McKendrick, A.G. (1926) Applications of Mathematics to Medical Problems. *Proc. Edinburgh Math. Soc.* 44, 98-103.

[52] Meng, X.L. (1994). Posterior Predictive p-values. *Ann. Statist.* 22, 1142-1160.

[53] Meng, X.-L. and Rubin, D.B. (1993) Maximum Likelihood Estimation via the ECM Algorithm: A General Gramework. *Biometrika.* 80, 267-278.

[54] Meng, X.-L. and van Dyk, D.A. (1997) The EM Algorithm - An Old Folk Song Sung to a Fast New Tune (with discussion). *J. Roy. Statis. Soc. B* 59, 511-567.

[55] Michiels, B., Molenberghs, G. and Lipsitz, S. R. (1999) Selection Models and Pattern-Mixture Models for Incomplete Data with Covariates. *Biometrics.* 55, 978-983.

[56] O'Malley, A. J. and Normand, S. L. (2005) Likelihood Methods for Treatment Noncompliance and Subsequent Nonresponse in Randomized Trials. *Biometrics.* 61, 325-334.

[57] Park, T. (1990) Estimation of the Nonresponse Models for Categorical Data. *Dissertation, Dept. of Biostat, Univ. of Michigan.*

[58] Peng, Y. H., Little, R. J. A. and Raghunathan, T. E. (2004) An Extended General Location Model for Causal Inferences from Data Subject to Noncompliance and Missing Values. *Biometrics.* 60, 598-607.

[59] Raghunathan, T. E. (2004) What Do We Do with Missing Data? Some Options for Analysis of Incomplete Data. *Annu. Rev. Public Health.* 25, 99-117.

[60] Robins, J.M., van der Vaart, A. and Ventura, V. (2000) Asymptotic Distribution of P Values in Composite Null Models (with discussion). *J. Am. Statist. Assoc.* 95, 1143-1172.

[61] Rosenheck, R., Cramer, J., Xu, W. C., Thomas, J., Henderson, W., Frisman, L., Fye, C. and Charney, D. (1997) A Comparison of Clozapine and Haloperidol in Hospitalized Patients with Refractory Schizophrenia. *New England Journal of Medicine.* 337, 809-815.

[62] Rubin, D. B. (1974) Characterizing the Estimation of Parameters in Incomplete Data Problems. *J. Am. Statist. Assoc.* 69, 467-474.

[63] Rubin, D. B. (1976a) Inference and Missing Data (with discussion). *Biometrika.* 63, 581-592.

[64] Rubin, D. B.(1978) Bayesian Inference for Causal Effects: the Role of Randomization. *The Ann. of Statis..* 6, 34-58.

[65] Rubin, D.B. (1981) Estimation in Parallel Randomized Experiments. *J. Educ. Statist.* 6, 377-400.

[66] Rubin, D.B. (1984) Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *Ann. Statist.* 12, 1151-1172.

[67] Rubin, D. B. (1987) Multiple Imputation for Nonresponse in Surveys. *New York: Wiley.*

[68] Rubin, D.B. (1996) Comments: On Posterior Predictive p-values. *Statistica Sinica.* 6 787-791.

[69] Rubin, D.B. and Schenker, N. (1986) Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *J. Am. Statist. Assoc.* 81, 366-374.

[70] Rubin, D.B., Stern, H. and Vehovar, V. (1996) Handling 'don't know' survey responses: the case of the Slovenian plebiscite. *J. Am. Statist. Assoc.* 90, 822-828.

[71] Schenker, N. (1985) Qualms About Bootstrap Confidence Intervals. *J. Amer. Statist. Assoc.* 80, 360-361.

[72] Sommer, A. and Zeger, S. (1991) On Estimating Efficacy from Clinical Trials. *Statis. Med.* 10, 45-52.

[73] Stubbendick, A. L. and Ibrahim, J.G. (2003) An AIC-type Model Selection Criterion for Missing Data Problems. *Technical Report, Dept. of Biostatistics, Univ. of North Carolina.*

[74] While, I. R. (2005) Uses and Limitations of Randomization-based Efficacy Estimators. *Statis. Methods in Medi. Res.* 14, 327-347.

[75] Woolson, R. F. and Clarke, W. R. (1984) Analysis of Categorical Incomplete Longitudinal Data. *J. Roy. Statist. Soc. A.* 147, 87-99.

[76] Zhou, X. H. and Li, S. M. (2006) ITT Analysis of Randomized Encouragement Design Studies with Missing Data *Statis. Med..* 25, 2737-2761.