REGRESSION ANALYSIS IN LONGITUDINAL STUDIES WITH NON-IGNORABLE MISSING OUTCOMES

by

Changyu Shen

B.S. in Biology

University of Science and Technology of China, 1998

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Biostatistics

University of Pittsburgh

2004

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Changyu Shen

It was defended on

March 22, 2004

and approved by

Lisa A. Weissfeld, Ph.D., Professor, Department of Biostatistics

Howard E. Rockette, Ph.D.,Professor, Department of Biostatistics

Sati Mazumdar, Ph.D., Professor, Department of Biostatistics

Gong Tang, Ph.D., Assistant Professor, Department of Biostatistics

Mary Ganguli, M.D., M.P.H., Professor, Department of Psychiatry and Department of Epidemiology

Hiroko H. Dodge, Ph.D., Assistant Professor, Department of Epidemiology

Dissertation Director: Lisa A. Weissfeld, Ph.D., Professor, Department of Biostatistics

REGRESSION ANALYSIS IN LONGITUDINAL STUDIES WITH NON-IGNORABLE MISSING OUTCOMES

Changyu Shen, PhD

University of Pittsburgh, 2004

One difficulty in regression analysis for longitudinal data is that the outcomes are often missing in a non-ignorable way (Little & Rubin, 1987). Likelihood based approaches to deal with non-ignorable missing outcomes can be divided into selection models and pattern mixture models based on the way the joint distribution of the outcome and the missing-data indicators is partitioned. One new approach from each of these two classes of models is proposed. In the first approach, a normal copula-based selection model is constructed to combine the distribution of the outcome of interest and that of the missing-data indicators given the covariates. Parameters in the model are estimated by a pseudo maximum likelihood method (Gong & Samaniego, 1981). In the second approach, a pseudo maximum likelihood method introduced by Gourieroux et al. (1984) is used to estimate the identifiable parameters in a pattern mixture model. This procedure provides consistent estimators when the mean structure is correctly specified for each pattern, with further information on the variance structure giving an efficient estimator. A Hausman type test (Hausman, 1978) of model misspecification is also developed for model simplification to improve efficiency. Separate simulations are carried out to assess the performance of the two approaches, followed by applications to real data sets from an epidemiological cohort study investigating dementia, including Alzheimer's disease.

PREFACE

This dissertation is organized as follows. In Chapter 1, I give an introduction on missing data problems in longitudinal data analysis. In Chapter 2, I review some approaches that have been used to analyze non-ignorable missing responses with some success and present the motivation of the two proposed methods described with greater detail in Chapter 3 and Chapter 4. Finally, I conclude this dissertation with a conclusion chapter.

I wish to thank my advisor, Dr. Lisa Weissfeld for her earnest help and guidance during the course of my dissertation work. It is her who introduced the concept of copulas to me, which later became one major component of my dissertation. Moreover, Dr. Weissfeld's continuous encouragement has greatly helped me build up my confidence in doing research and her insightful suggestions clarified lots of problems that had confused me. I truly am grateful to Dr. Weissfeld's patient revision of my proposal and this dissertation.

I would like to thank Dr. Howard E. Rockette, Dr. Sati Mazumdar, Dr. Gong Tang, Dr. Mary Ganguli and Dr. Hiroko H. Dodge for serving as my committee members. I want to thank Dr. Gong Tang for valuable discussions that have substantially deepened my understanding of missing-data problems. I want to thank Dr. Howard E. Rockette and Dr. Sati Mazumdar, whose lectures on estimation theory and linear models established indispensable background for me to accomplish the application of pseudo maximum likelihood method on pattern mixture models.

Special thanks to Dr. Mary Ganguli for providing me an opportunity to practise statistics in psychiatric research area and continuous financial support. Special thanks to my GSR supervisor, Dr. Hiroko Dodge, who taught me a lot on applied statistics throughout years of collaboration and inspired me with a semi-parametric pattern mixture model.

Finally, I dedicate this dissertation to my parents and sister for their love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

1.1 MISSING DATA IN LONGITUDINAL STUDIES

Longitudinal studies are characterized by a series of measurements of interest on the same unit over time. Failure to obtain a full set of observations results in incomplete data or missing data. This phenomenon is quite common in longitudinal studies. For example, in an epidemiological cohort study, subjects might drop-out, move away, be too sick to measure data on, die, and so forth. This can result in a serious missing-data problem. In general, the missing-data pattern can be divided into either a monotone missing or an intermittent missing-data pattern. The first one refers to the scenario where all observations after a certain follow-up time are missing, so that all other missing scenarios belong to the intermittent missing-data pattern.

One point that needs to be emphasized is that, in most longitudinal studies, knowledge of the mechanisms that lead to certain values being missing is usually very limited. Missing-data mechanisms can be very complicated and diversified due to the various reasons that cause missing values. This is quite different from other type of studies, in which the missing-data process is under the control of the individuals who design the experiment. For example, double sampling is a survey methodology where missing data arise as part of the sampling scheme. A sample is selected from the population and some characteristics are measured. Then a subsample is selected from the original one and more variables are measured. Thus the missing-data process is related to the extra variables that are not measured in each subject in the original sample and is under the control of the sampler, who has all the knowledge pertaining to the missing-data mechanism. Therefore, missing data in longitudinal studies pose a much more difficult problem for statistical inference.

The question of greatest interest is related to the handling of missing data in an analysis. That is, if we analyze the observed data and ignore the missing-data process, what and how much do we lose, as compared with what we would obtain were there no missing data? It is obvious that we lose efficiency since there are few data available for estimation and inference. For example, suppose we want to estimate the mean of a variable from a population. If we randomly lose 1/3 of the data, then the variance of the observed sample mean is 50% larger than the variance of the complete sample mean. In fact, a more serious problem in most longitudinal analyses is that we might have a biased estimate. Although this is not always the case, it can be enough of a problem to result in an invalid analysis. In the example above, instead of random missingness, suppose that all of the measurements smaller than a certain value are lost due to some unknown reason. Then the mean of the observed measurements is obviously invalid or biased since the observed data are actually from a truncated distribution of the underlying one. Therefore, the key problem is the form of the missing-data mechanism, on which we often have little, if any, information.

Clearly, the handling of missing data is not only a statistical issue, but also related to the process of data collection and the scientific problem at hand. Statisticians have been trying to develop strategies and methods to reduce the bias and improve the efficiency for the last two decades. However, it should be emphasized that in order to obtain a valid inference when missing data are present, one should gather any information regarding the data collection and the missing-data process. The optimal strategy is of course to avoid missing data by every effort. However, it is often not possible to do this.

One naive way to deal with multivariate data with missing values is to ignore the observations with missing values and analyze only the complete cases (complete case analysis or CC). This is obviously not efficient and can seriously bias the results of an analysis. Apart from the CC approach, current statistical procedures to deal with missing values can be divided into three classes based on the nature of the approaches: (i) imputation-based procedures; (ii) weighting procedures and (iii) model-based procedures.

(i) Imputation-based procedures

Basically, this type of procedure fills in the blanks in the data set with some values based on the observed data to create a full data set. Then standard procedures are applied on

the "full" data set. Essentially, the full data set does not have more information than the observed data can provide us since the imputation is based on the observed data. This procedure has been used extensively in many epidemiological studies. Two typical imputation methods are mean imputation and conditional mean imputation. The first one fills in the blanks with the means of observed values. For example, a missing measurement of a certain variable at a certain time point is taken to be the mean of the two observed measurements that are closest to it on the time axis. Conditional mean imputation fills in the blanks by the predictive values based on a regression model fit to the observed data. For example, suppose some responses are not observed in a regression analysis. Then one can fit a regression model on the observed response and predict the unobserved response based on the corresponding covariates (assume that covariates are always observed). Currently, a multiple imputation technique (Rubin, 1987) is a popular tool for many missing-data problems. Essentially, this approach creates m complete data sets based on the observed data and some model and then standard statistical procedures are applied to each one of them. Finally, the results from the m data sets are combined to make the ultimate inference. This technique is "designed to handle the problem of missing data in public-use data bases where the data-base constructor and the ultimate user are distinct entities and there typically is no one accepted reason for the missing data" (Rubin, 1996).

(ii) Weighting procedures

The idea here is to analyze the observed data by assigning different weights to different observations (Robins et al., 1995). The full data can be seen as a random sample from the underlying distribution and hence each observation is equally weighted. The missing-data process causes some observations to be observed with greater likelihood than others. Thus the intention is to put more weight on those observations that are less likely to be observed. This type of approach is mostly used in analyzing sample survey data where the weights are known due to the sampling scheme. However, the concept is quite general and later we will see that some model-based approaches share the same strategy.

(iii) Model-based procedures

Most of the current regression analysis techniques for missing data belong to this class of procedures. A model is built for the data, including the missing-data process, and inference

is drawn directly from the model. The model based procedures are particularly useful to handle responses that are Missing Not At Random (MNAR). This class of procedures will be discussed in greater detail in Section 1.2.

In summary, missing data are very common in longitudinal studies. The missing-data process can be very complicated and ignoring it can sometimes lead to invalid inference. The three major statistical approaches discussed here have been used to analyze data sets with missing values with some success. However, the model-based procedures and newer approaches based on weighting offer the most flexible approaches to attacking these problems.

1.2 MODEL-BASED APPROACHES

Longitudinal studies record many variables on the same experimental unit repeatedly. Usually regression analysis is used to relate a variable to other variables in the data set. To establish the notation, the subject index is suppressed throughout this section for simplicity. Let $Y = (Y_1, Y_2, \ldots, Y_m)^T$ be the response vector and $X = (X_1^T, X_2^T, \ldots, X_m^T)^T$ be the covariate matrix, where X_i is a row vector coding the covariates at time i. Therefore (X, Y) is the full data. In the following context, it is assumed that X is always fully observed and elements of Y are subject to missingness. There are three types of missing-data mechanisms that are defined as follows (Rubin, 1976; Little, 1995):

(i) Missing Completely At Random (MCAR): the probability that y_i is observed depends neither on X nor Y. Under this assumption, the observed response y_i can be treated as a random sample from the distribution of Y_i for $i = 1, 2, \ldots, m$.

(ii) Missing At Random (MAR): the probability that y_i is observed can depend on the observed y values and X as well, but not on unobserved y values. Therefore MCAR is just a subclass of MAR. Another subclass of MAR that is bigger than MCAR is the Covariate Dependent Missing (CDM). In that case, the probability of being observed can depend only on X. Under CDM, the observed response y_i is not necessarily a random sample from the distribution of Y_i, it is a random sample from the conditional distribution $[Y_i|X]$.

(iii) Missing Not At Random (MNAR) or Non-Ignorable Missing (NIM): if the missing-data process is not MAR, it usually belongs to this class. Basically, it says that

the probability that y_i is observed depends on unobserved y values (possibly on observed y values and X as well). Since whether or not a missing-data process is ignorable depends on what approach one takes for the analysis, "ignorable" here means that the missing-data mechanism can be ignored when using likelihood based approaches to analyze the data.

When a missing-data process is MCAR or MAR and the parameters associated with the missing-data process are distinct from the regression parameters, likelihood based approaches ignoring the missing-data process can be applied to make valid inference. However, in many longitudinal studies, the missing-data process involved belongs to the MNAR class, which turns out to be the most difficult missing-data mechanism to deal with since unverifiable assumptions regarding the missing-data process have to be made for inference. In this case, model based approaches have to take into account the missing-data mechanism in order to draw valid inferences about the distribution of the response.

To write out the likelihood function for the model based approach, let $y = (y_{obs}^T, y_{mis}^T)^T$ be a full realization, where y_{obs} and y_{mis} are the observed part and missing part of y, respectively. Furthermore, let $R = (R_1, R_2, \ldots, R_m)^T$ be a vector composed of binary variables such that $R_i = 1$ indicates that y_i is observed and 0 otherwise. Therefore the observed likelihood is obtained by integrating out the missing values:

$$L(\lambda; y_{obs}, r) = \int p(y_{obs}, y_{mis}, r|X, \lambda) dy_{mis}, \tag{1.1}$$

where λ is the parameter that indexes the conditional distribution of (Y, R) given X. There are two ways to partition this distribution, which results in two classes of models that tackle the missing problem differently. In the following discussion, it is always assumed that X is fixed. Thus any distribution that is mentioned is actually the distribution conditional on X.

(i) Selection models

This class of models partitions the joint distribution of (Y, R) to be the distribution of Y and the conditional distribution of R given Y. Therefore,

$$p(y, r|X, \lambda) = p(y|X, \beta)p(r|y, X, \gamma), \tag{1.2}$$

where $\lambda = (\beta, \gamma)$ and β indexes the distribution of Y and γ indexes the conditional distribution of R given Y. In most practical studies β and γ are assumed to be distinct. The

terminology "selection model" is originally from the sample survey literature. In that area, $p(r|y, X, \lambda)$ represents the probability of being "selected" given the characteristics. Then if the missing-data process is MAR, or $p(r|y, X, \gamma) = p(r|y_{obs}, X, \gamma)$, (1.1) can be written as

$$\begin{aligned} L(\beta, \gamma; y_{obs}, r) &= \int p(y_{obs}, y_{mis}, r|X, \lambda) dy_{mis} \\ &= \int p(y_{obs}, y_{mis}|X, \beta) p(r|y_{obs}, X, \gamma) dy_{mis} \\ &= p(y_{obs}|X, \beta) p(r|y_{obs}, X, \gamma). \end{aligned}$$

Thus the likelihood function ignoring the missing-data process is proportional to the full likelihood function when β and γ are assumed to be distinct. Therefore we can just ignore the missing-data process and apply the standard likelihood based methods (e.g. MLE) to the observed data.

Because of the problems that arise with missing data it is natural to ask if we can test if the data is MAR or MNAR. If the test does not reject MAR, then we can ignore the missing-data process and proceed with standard methods. Unfortunately, it is impossible to construct this type of test based on the observed data (Laird, 1988). Or in other words, the observed data provide insufficient information about $p(r|y, X, \gamma)$. In this sense, selection models are under-identified.

(ii) Pattern-mixture models

As an alternative to selection models, pattern-mixture models partition the joint distribution of (Y, R) to be the distribution of R and the conditional distribution of Y given R, i.e.

$$p(y, r|X, \lambda) = p(r|X, \pi) p(y|r, X, \phi), \tag{1.3}$$

where $\lambda = (\pi, \phi)$ and π indexes the distribution of R and ϕ indexes the conditional distribution of Y given R. Thus the observed likelihood function can be written as

$$\begin{aligned} L(\pi, \phi; y_{obs}, r) &= \int p(r|X, \pi) p(y_{obs}, y_{mis}|X, r, \phi) dy_{mis} \\ &= p(r|X, \pi) p(y_{obs}|X, r, \phi). \end{aligned}$$

Therefore if π and ϕ are distinct, they can be estimated separately. Moreover, one can build entirely different models for the response given a certain missing-data pattern.

From the construction of pattern-mixture models one can see that not all of the parameters associated with the model are estimable (e.g. not every element of ϕ is estimable). Obviously, the data provide no information on the regression parameters of the unobserved response as a function of the observed response for any missing-data pattern. Consider bivariate data (Y_{i1}, Y_{i2}) with Y_{i1} observed for all i and Y_{i2} subject to missingness. Therefore the data are composed of two patterns: both variables are observed and only Y_{i1} is observed. Assume that for each of the two patterns, the two variables follow a bivariate normal distribution. Then there is no way we can estimate the regression parameters of Y_2 on Y_1 for the incomplete pattern. Or in other words, the observed data provide no information on the marginal distribution of the second variable and the correlation between the two variables for the incomplete pattern. Therefore, pattern-mixture models are also under-identified. In order to fully identify the model, one needs to put certain constraints on the parameter space of the model. This issue will be discussed with more detail in Chapter 2.

Selection models partition the joint distribution of the response and the missing-data indicators in a natural way. Therefore it has been popular for the analysis of missing data. However, to evaluate the likelihood function, one needs to integrate out the missing values. Although the EM algorithm and Markov Chain Monte Carlo (MCMC) procedures provide powerful tools to optimize complicated integrals, the computational burden inevitably limits its practical application. Moreover, selection models can be very sensitive to misspecification of the missing-data process. If the selection process we propose deviates from the true missing-data process in a certain manner, then the estimator can be very biased. Pattern-mixture models formulate different models for different missing-data patterns and standard statistical procedures are applied to each pattern. Therefore the computation is straight forward and one does not have to specify the missing-data process. However, constraints on the parameter space can be very hard to postulate. Moreover, often one is interested in the marginal distribution of the response, which is a weighted sum of the distribution within each pattern. Thus pattern-mixture models do not model the marginal distribution of the response in a direct way.

2.0 MODEL-BASED APPROACHES TO THE ANALYSIS OF MISSING DATA

In this chapter, methods that haven been used to analyze longitudinal data with non-ignorable missing responses in the statistical literature are briefly reviewed, followed by the motivation of the two proposed methods described in Chapter 3 and Chapter 4.

2.1 SELECTION MODELS

Wu & Carroll (1988) developed a likelihood-ratio test for informativeness of a monotone missing-data process and maximum likelihood estimates for the expected rates of change and the parameters of the right-censoring process under the framework of a linear random-effects model with a probit model for the right-censoring process. In their model, the missing-data process can depend on the random effects and thus depends on the unobserved response indirectly. When such a dependency does exist, or when the missing-data process is informative, they showed that the bias can be substantial if the dependency is ignored.

Diggle & Kenward (1994) used a similar strategy to analyze monotone missing data. They assumed a linear Gaussian model for the response and a logistic model for the missing-data process. In the logistic model, the probability of dropping out of the study at a certain time point is assumed to be dependent on the current and the previous responses, but not on the future responses. They used the simplex algorithm by Nelder & Mead (1965) to maximize the observed likelihood function. Evaluation of the likelihood function requires numerical integration, which was carried out by the probit approximation to the logit transformation. Troxel, Harrington & Lipsitz (1998a) extended Diggle & Kenward's (1994) approach to in-

clude non-ignorable non-monotone missing data. Again, the authors assumed a multivariate normal model for the response and a logistic model for the missing-data indicator given the current response (perhaps the previous responses as well). To fit within the framework of the non-monotone missing-data pattern, they also assumed a first order Markov property for the multivariate normal model. The two papers introduced above clearly demonstrate that the estimation task in a selection model may require a substantial amount of computation due to the complexity of the likelihood function.

Follmann & Wu (1995) proposed a "shared parameter" model to deal with missing outcomes in longitudinal data. They assumed separate models for the response and the missing-data process, which are linked by a common random parameter. Conditional on the random parameter, they assumed independent generalized linear models for the response and the missing-data process. Based on this setting, the response and the missing-data process are connected in an implicit way, which as shown by the authors, includes the non-ignorable missing-data mechanism. Although the explicit observed likelihood function can be written out as an integral over the range of the random parameter, it is rather complicated. Therefore the authors based their inference on the distribution of the response given the number of observed responses. With a simulation study, they demonstrated that the empirical Bayesian estimates based on the conditional distribution can recover the true values.

Troxel, Lipsitz & Harrington (1998b) proposed a marginal approach for longitudinal measurements with non-ignorable non-monotone missing data. They assumed a normal model for the marginal distribution and a logistic model for the missing-data indicator of the current response given the current response. The authors further forced the outcomes and the associated missing-data indicators to be independent within each subject. Estimates of parameters were calculated by maximizing the "pseudolikelihood function". It was shown that the estimators are consistent and asymptotically normal, provided that the model for the missing-data indicator given the corresponding response was correctly specified. However, the authors did emphasize that the estimates can be sensitive to misspecification of the logistic model. Ibrahim, Chen & Lipsitz (2001) proposed an approach for a non-ignorable non-monotone missing response in the framework of a generalized linear mixed model. In modelling the missing-data process, they partitioned the joint distribution of the missing-

data indicators into a series of conditional distributions with each conditional distribution assumed to follow a logistic model. Under the assumption that the missing-data indicator depends only on the response and the possible covariates, but not on the random effects, they developed a Monte Carlo EM algorithm to maximize the observed likelihood function.

Robins et al. (1994, 1995) proposed a class of semi-parametric estimators for repeated outcomes in the presence of missing data in the framework of weighted estimating equations. Their model for the response only requires correct specification of the functional form of the mean. They assumed that the missing-data process is either known or can be estimated based on a parametric model. They showed that their estimators are consistent and that the asymptotic variance of the optimal estimator in their class reaches the semi-parametric variance bound. Other selection model based research includes Conaway (1992, 1993), Little (1995) and Robins (1997).

2.2 PATTERN-MIXTURE MODELS

Little (1993) discussed pattern-mixture models for multivariate data with a general pattern for the missing values, with an emphasis on the restrictions of non-identifiable parameters. In the paper, the author defined the missing variable (MV) distribution for each missing-data pattern to be the conditional distribution of the missing values given the observed values. One simple restriction is the complete-case missing variable (CCMV) restriction, which equates all parameters associated with MV distributions to their identifiable analogs in the stratum of complete cases. Consider the example of bivariate data with the second variable subject to missingness. We again assume that for each pattern, the data follow a bivariate normal distribution. If we let $\phi^{(r)} = (\mu_1^{(r)}, \mu_2^{(r)}, \sigma_{11}^{(r)}, \sigma_{22}^{(r)}, \sigma_{12}^{(r)})$ denote the means, variances and covariance of Y_1 and Y_2 for pattern r, where $r = 1$ refers to the complete case and $r = 2$ refers to the pattern with Y_2 missing. Then obviously $(\mu_2^{(2)}, \sigma_{22}^{(2)}, \sigma_{12}^{(2)})$ are not identifiable. By reparameterization, it is equivalent to the non-identifiability of $(\beta_{0\cdot 12}^{(2)}, \beta_{1\cdot 12}^{(2)}, v_{12}^{(2)})$, where $\beta_{0\cdot 12}^{(2)}$, $\beta_{1\cdot 12}^{(2)}$ and $v_{12}^{(2)}$ are the intercept, coefficient and variance of the error term in the regression of Y_2 on Y_1 for $r = 2$. In this case, the CCMV restriction will equate these regression parameters to their analogs in the complete cases: $(\beta_{0\cdot 12}^{(2)}, \beta_{1\cdot 12}^{(2)}, v_{12}^{(2)}) = (\beta_{0\cdot 12}^{(1)}, \beta_{1\cdot 12}^{(1)}, v_{12}^{(1)})$.

An alternative to the complete case restriction is to equate the non-identifiable parameters in a missing-data pattern to a weighted sum of the identifiable counterparts in a set of other patterns. A natural choice for the weight is the proportion of each pattern in the set.

Little (1994) discussed a pattern-mixture model for bivariate monotone data, assuming that the response follows a bivariate normal distribution for each of the two patterns. In the paper, the author demonstrated the nature of the restriction that is embedded within a pattern-mixture model under different assumptions for the missing-data process. In particular, when the probability of being a complete case depends only on the first or the second variable, a complete case restriction is automatically embedded within the pattern-mixture model. Large sample inference and small sample Bayesian inference were also described in the paper. Little & Wang (1996) further extended this approach to multivariate incomplete data with covariates. The data they analyzed were of a special format: certain elements of the outcome vector were always observed and other elements were either fully observed or were always missing. Thus the data can be seen as a general format of the data analyzed in Little (1994). Maximum likelihood and Bayesian methods were used to estimate regression parameters.

Park & Lee (1999) applied a pattern-mixture model to urinary incontinence data within the framework of generalized estimating equations. Subjects have at most three repeated binary measurements with the last two subject to missingness, leading to four missing-data patterns in the data. The CCMV assumption was used for the scale and correlation parameters. Estimates of the regression parameters were obtained by plugging into the GEE the pattern indicators and their interactions with other covariates, with the non-identifiable parameters set to be equal to their analogs in the complete cases. Kenward, Molenberghs & Thijs (2003) identified a family of restrictions where drop-out does not depend on future unobserved observations for pattern-mixture models.

The major difficulty associated with pattern-mixture models is the under-identifiability of such models. Two major strategies to deal with this problem are identifying restrictions and model simplification (Thijs et al., 2002). For the first strategy, non-identifiable MV distributions of the incomplete patterns are set equal to functions of their identifiable analogs in other patterns (e.g. the CCMV restriction). On the other hand, the second strategy allows

different patterns to share certain parameters so that the incomplete patterns can borrow information from patterns with more data points. Roy (in press) proposed a slightly different approach within the framework of pattern-mixture models based on latent class variables, which is similar to the second strategy described above.

2.3 MOTIVATION AND CONTRIBUTION OF THE PROPOSED METHODS

2.3.1 NORMAL COPULA BASED SELECTION MODELS

The key feature of a selection model is the modelling of the missing-data process. To simplify the likelihood function, one might want to specify certain restrictions on the distribution of the response (e.g. Markov property) and the missing-data process (e.g. dependent on the history of the response but not the future response). However, estimates obtained from selection models can be sensitive to misspecification of the missing-data mechanism. Therefore, as we simplify the problem, we are paying a price for a possible serious deviation from the true values. Ideally, one wants to include as many mechanisms as possible in their selection model in order to include the true mechanism or a mechanism that is close to the true one. This inevitably increases the complexity of the inference. One strategy is to find a balance point between these two extremes. For the modelling of multivariate outcome data, the use of the copula function has proved to be a promising approach (Nelsen, 1998). Hence, it can be used to model the correlation between the response and its missing-data indicators. A copula family can provide a diversified correlation structure, yet maintain the simplicity of the model to a reasonable extent.

In Chapter 3, a selection model based on a normal copula is proposed in the hope of balancing simplicity and robustness. This method models the joint distribution of a continuous outcome and its missing-data indicators directly through a multivariate normal copula function, which, as a result, defines a class of missing-data mechanism. The method allows one to check the correlation between the outcome and the missing-data indicators by the estimates of the parameters associated with the copula function. Since a copula function

defines a correlation structure among random variables that is independent of their margins, it also allows one to experiment different margins and copula families.

2.3.2 PATTERN-MIXTURE MODEL WITH PSEUDO MAXIMUM LIKELIHOOD ESTIMATION

For pattern-mixture models, there are two issues that have not received sufficient attention. First, the conditional distribution of the response given a missing-data pattern can be very complicated even though the marginal distribution is of a simple form, mainly due to the complexity of the missing-data process. Therefore a semi-parametric model without the need to specify the distributional form for each pattern is desirable. Second, as Hogan and Laird (1997) have pointed out, the dimension of the parameter vector in general pattern-mixture models is so large that estimation requires that each missing-data pattern occurs sufficiently often. Therefore, model simplification has been an important issue for this class of models. Currently, most proposals have been focused on allowing incomplete patterns and patterns with more data points to share certain parameters so that the model is identifiable. For instance, in a model for a continuous outcome where time is the only covariate and the missing-data process is monotone, the slope of time is inestimable for patterns with only 1 observation. Then setting it equal to the slope for the pattern with two observations makes the model identifiable if the functional form of time is linear. However, no one has focused on further simplification of the identifiable model in a systematic manner. Thus the focus of this work is to create models where patterns share parameters in order to maintain identifiability and improve efficiency. Considering the above example, one might want to ask if the intercept is different across patterns or some patterns share the same intercept? This is an important supplement to the issue of under-identifiability since one already loses some efficiency by missing data.

In Chapter 4, a semi-parametric approach armed with a model misspecification test is presented to address these two issues. This method only requires the specification of the first (and second) moment(s) of each pattern for estimation since often we are unwilling to make assumptions about the distributional form due to the complexity of the missing-data

process. The estimation procedure is a direct generalization of the Pseudo Maximum Likelihood Estimation proposed by Gourieroux et al. (1984), which is itself a multi-dimensional extension of the Quasi Likelihood Function discussed by Wedderburn (1974). The model misspecification test statistic is readily calculated after the estimates are obtained. The test provides clues on model simplification, which captures the major differences across patterns and ignore the small ones in order to improve efficiency.

The proposed approach is similar to that of Park & Lee (1999) and Fitzmaurice & Laird (2000). However, none of the above authors raised the issue of efficiency improvement. Moreover, the new treatment on the methodology provides a more rigorous and general framework to simplify identifiable pattern-mixture models.

3.0 NORMAL COPULA BASED SELECTION MODELS

In this chapter, a class of selection models based on a multivariate normal copula function is introduced. It turns out that the normal copula specifies a class of selection models that can be very useful for the modelling of continuous missing data. The conditional distribution of the missing-data indicators R given Y_{obs} is expressed as a weighted sum (with weight being 1 or -1) of a multivariate normal CDF evaluated at different points. This is the extra term that is used to adjust the likelihood function due to non-ignorable missingness. This approach is particularly useful when the number of repeated measures is not large and can be applied to non-monotone missing scenarios.

The major advantage of copula-based selection models is that the correlation structure between the outcome and its missing-data indicators and their marginal distributions can be modelled independently of each other. This allows for flexibility in the choice of both the margins and the copula that defines the missing-data process. Many selection models are constructed through a "shared parameter", conditional on which the outcome and the missing-data indicators are independent. Although this hierarchical structure allows one to fit complicated models, it is often hard to see how the outcome and the missing-data process are correlated. On the other hand, the normal copula based selection models allow one to check the correlation directly from the estimates of the parameters associated with the copula function.

Selection models require a substantial amount of computation due to the complicated integrals to be evaluated. In the proposed selection models, this is reduced to the evaluation of the CDF of a multivariate normal distribution at different points, which can be approximated by the Monte Carlo approach since sampling from a multivariate normal distribution has been standardized in many software packages. Another advantage of the copula-based selec-

tion models is that a pseudo maximum likelihood estimation procedure (Gong & Samaniego, 1981) can be readily applied to reduce the dimension of the parameter vector for optimization.

The concept of copula functions and the multivariate normal copula function are described in Sections 3.1 and 3.2, respectively. The general form of the model is presented in Section 3.3. In Section 3.4, the robustness of the model under different missing-data processes that differ from that specified by the copula model is studied through simulations. The proposed method is applied to a real data set from an epidemiologic study in Section 3.5, followed by a discussion section.

3.1 COPULA

A copula function can be thought as a link function between a multivariate distribution function and its marginal univariate distributions (Nelsen, 1998). In this section, the bivariate copula functions are first introduced, followed by an extension to multivariate copula functions.

3.1.1 BIVARIATE COPULA FUNCTIONS

Some preliminaries are needed to lead to the mathematical definition of a bivariate copula function. As will be seen from the process of building a copula function, a function has to satisfy certain conditions to be the link function of a bivariate distribution and its marginal distributions. Nevertheless, many families of bivariate copula functions have been identified and they provide diversified dependence structures between two random variables.

Let $\mathbf{R}$ denote the real line $(-\infty, \infty)$, $\bar{\mathbf{R}}$ denote the extended real line $[-\infty, \infty]$, and $\bar{\mathbf{R}}^2$ denote the extended real plane. A *rectangle* in $\bar{\mathbf{R}}^2$ is the Cartesian product B of two closed intervals: $B = [x_1, x_2] \times [y_1, y_2]$. The *vertices* of a rectangle B are the points $(x_1, y_1), (x_1, y_2), (x_2, y_1)$ and (x_2, y_2). The *unit square* $\mathbf{I}^2$ is the product $\mathbf{I} \times \mathbf{I}$ where $\mathbf{I} = [0, 1]$. Then a *2-place real function* H is a function whose domain, Dom$H$, is a subset of $\bar{\mathbf{R}}^2$ and whose range, Ran$H$, is a subset of $\mathbf{R}$.

Let S_1 and S_2 be nonempty subsets of $\bar{\mathbf{R}}$, and let H be a function such that Dom$H$=$S_1 \times S_2$. Let $B = [x_1, x_2] \times [y_1, y_2]$ be a rectangle all of whose vertices are in Dom$H$. Then the *H-volume* of B is given by

$$V_H(B) = H(x_2, y_2) - H(x_2, y_1) - H(x_1, y_2) + H(x_1, y_1).$$

Then a 2-place real function H is *2-increasing* if $V_H(B) \geq 0$ for all rectangles B whose vertices lie in Dom$H$.

Suppose S_1 has a least element a_1 and that S_2 has a least point a_2. We say that a function H from $S_1 \times S_2$ into $\mathbf{R}$ is *grounded* if $H(x, a_2) = 0 = H(a_1, y)$ for all $(x, y) \in S_1 \times S_2$. It can be shown that a grounded 2-increasing function with domain $S_1 \times S_2$ is nondecreasing in each of its two arguments.

Now we are ready to define a bivariate copula function. A *bivariate copula function* is a function C with the following properties:

(i) Dom$H$=$\mathbf{I}^2$;

(ii) C is grounded and 2-increasing;

(iii) For every u and v in $\mathbf{I}$,

$$C(u, 1) = u \text{ and } C(1, v) = v.$$

Note that for every (u, v) in $\mathbf{I}^2$, $0 \leq C(u, v) \leq 1$.

Up to this point, we have not mentioned any roles that a bivariate copula function might play in the relationship between bivariate distribution functions and their univariate margins. This is elucidated by Sklar's theorem, which has been the foundation of many applications of copula functions.

Sklar's Theorem: *Let H be a bivariate joint distribution function with margins F and G. Then there exists a copula C, such that for all x and y in $\bar{\boldsymbol{R}}$,*

$$H(x, y) = C(F(x), G(y)).$$

If F and G are continuous, then C is unique; otherwise, C is uniquely determined on $\mathrm{Ran}F \times \mathrm{Ran}G$. *Conversely, if C is a copula and F and G are distribution functions, then the function H defined above is a joint distribution function with margins F and G.*

Thus a bivariate copula function can be thought of as a bivariate cumulative distribution function with margins that are uniform on $\mathbf{I}$. The significance lies in that bivariate copula functions provide a way to construct different bivariate distribution functions based on fixed margins. On the other hand, it is also possible for different bivariate distributions to share the same copula function.

3.1.2 MULTIVARIATE COPULA FUNCTIONS

The idea of bivariate copula functions extends to the multivariate case quite naturally. Let $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ be two points in $\bar{\mathbf{R}}^n$. We say $\mathbf{a} \leq \mathbf{b}$ if $a_k \leq b_k$ for all k; and $\mathbf{a} < \mathbf{b}$ if $a_k < b_k$ for all k. For $\mathbf{a} \leq \mathbf{b}$, we will use $[\mathbf{a}, \mathbf{b}]$ to denote the *n-box* $B = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$. The *vertices* of an n-box B are the points $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ where each c_k is equal to either a_k or b_k. An *n-place real function* H is a function whose domain, $\mathrm{Dom}H$, is a subset of $\bar{\mathbf{R}}^n$ and whose range, $\mathrm{Ran}H$, is a subset of $\mathbf{R}$.

Suppose $S_1, S_2, \ldots, S_n$ are nonempty subsets of $\mathbf{R}$ and H is an n-place real function such that $\mathrm{Dom}H = S_1 \times S_2 \times \cdots \times S_n$. Let $B = [\mathbf{a}, \mathbf{b}]$ be an n-box all of whose vertices are in $\mathrm{Dom}H$. Then the *H-volume of* B is given by

$$V_H(B) = \sum_{\mathbf{c}} \mathrm{sgn}(\mathbf{c}) H(\mathbf{c}),$$

where the sum is taken over all vertices $\mathbf{c}$ of B, and $\mathrm{sgn}(\mathbf{c})$ is given by

$$\mathrm{sgn}(\mathbf{c}) = \begin{cases} 1, & \text{if } c_k = a_k \text{ for an even number of } k\text{'s}, \\ -1, & \text{if } c_k = a_k \text{ for an odd number of } k\text{'s}. \end{cases}$$

Then an n-place real function H is *n-increasing* if $V_H(B) \geq 0$ for all n-boxes B whose vertices lie in $\mathrm{Dom}H$.

Now suppose that each S_k has a least element a_k. We say that H is *grounded* if $H(\mathbf{t}) = 0$ for all $\mathbf{t}$ in $\mathrm{Dom}H$ such that $t_k = a_k$ for at least one k. It can be shown that a grounded

n-increasing function with domain $S_1 \times S_2 \times \cdots \times S_n$ is nondecreasing in each argument. Furthermore, if each S_k has a greatest element b_k, then we say that H has *margins*, and the *one-dimensional margins* of H are the functions H_k given by $\mathrm{Dom}H_k = S_k$ and

$$H_k(x) = H(b_1, \ldots, b_{k-1}, x, b_{k+1}, \ldots, b_n).$$

Thus an *n-dimensional copula* is a function C with the following properties:
(i) $\mathrm{Dom}C = \mathbf{I}^n$;
(ii) C is grounded and n-increasing;
(iii) Each margin C_k $(k = 1, 2, \ldots, n)$ satisfies

$$C_k(u) = u \text{ for all } u \text{ in } \mathbf{I}.$$

Note that for every $\mathbf{u}$ in $\mathrm{Dom}C$, $0 \leq C(\mathbf{u}) \leq 1$. Moreover, the 2-dimensional Sklar's theorem extends to the n-dimensional case directly.

Copula functions provide structural information about multivariate distributions. On one side, the marginal distributions determine the behavior of each component as a single random variable. The copula function then determines how these random variables behave as a random vector. In other words, the copula function reveals the correlation structure embedded within a multivariate distribution. It is this property of copula functions that makes it a useful tool for the modelling of correlated repeated measurements. There are many copula families that represent different correlation structures and hence provide for a wide range of models. For the approach to be discussed later in this chapter, a normal copula function is used to model continuous outcomes and their missing-data indicators simultaneously. The hope is that the proposed normal copula family can capture the essential dependence between the outcomes and the missing-data indicators in order to reduce the bias that would arise if we ignore the missingness.

3.2 NORMAL COPULA

One example of a multivariate copula function is the multivariate normal copula, which is of the form:

$$C_\Omega(u_1, u_2, \ldots, u_m) = \Phi(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_m)|0, \Omega),$$

where Ω is a correlation matrix, $\Phi(.)$ denotes the standard normal CDF and $\Phi(.|0, \Omega)$ denotes the CDF of a multivariate normal vector $(m \times 1)$ with mean 0 and covariance matrix Ω. It is obvious that any subset of the m variables also has a multivariate normal copula whose correlation matrix is a submatrix of Ω that matches the corresponding elements in the subset. Clearly, Ω provides the essential correlation structure among the m variables, which is free of the margins.

For a multivariate CDF with a normal copula function, Ω does not necessarily stand for the correlation matrix for the random vector unless the margins are normal distributions. The question is how Ω affects the dependence among the elements of the random vector. It turns out that this type of "dependency" is closely connected to Kendall's tau . Let (X_1, Y_1) and (X_2, Y_2) be i.i.d. continuous random bivariate vectors with joint distribution function H. Then Kendall's tau is defined as the probability of concordance minus the probability of discordance:

$$\tau_{X,Y} = \Pr[(X_1 - X_2)(Y_1 - Y_2) > 0] - \Pr[(X_1 - X_2)(Y_1 - Y_2) < 0].$$

It is easy to see that the range of Kendall's tau is [-1,1]. For a bivariate random vector, Kendall's tau is a measure of the likelihood that "large" values of one variable tend to result in "large" values of the other variable. It is precisely the copula function that captures this kind of dependence. In fact, if X and Y are continuous random variables whose copula is C, then Kendall's tau is given by

$$\tau_{X,Y} = 4 \int \int_{\mathbf{I}^2} C(u, v) dC(u, v) - 1.$$

In figure 3.1, the relationship between ρ and τ is shown for a bivariate normal copula function, where ρ is the off-diagonal element of Ω. It is clear from the graph that ρ has an approximate linear relationship with τ with a positive slope within the range of (-0.5, 0.5). Graphically, ρ can be seen as a new version of τ that stretches τ towards -1 and 1. Although ρ need not be the correlation coefficient, it reflects the dependence of the two variables in a way that is similar to Kendall's τ. Therefore, ρ can be thought as a measure of dependence

between two random variables. Note that this measure is independent of the marginal distributions of the two random variables.

There are two major advantages to using a normal copula function to model the joint distribution of the response and missing-data indicators: (i) the elements of Ω represent the dependence of corresponding variables and (ii) the margins can change freely. An additional advantage of using a normal copula to build selection models is that one can apply the Monte Carlo method directly to avoid numerical integration since sampling from a multivariate normal distribution has been standardized in many statistical software packages.

3.3 MODEL SPECIFICATION AND ESTIMATION

3.3.1 MODEL SPECIFICATION

To ease the notation, the index denoting subject is suppressed unless otherwise noted. Let $Y = (Y_1, Y_2, \ldots, Y_m)^T$ be an $m \times 1$ continuous outcome vector and X be an $m \times p$ covariate matrix. Let X_j be the jth row of X and $R = (R_1, R_2, \ldots, R_m)^T$ be the missing-data indicator vector. Here $R_j = 1$ indicates that y_j is observed and $R_j = 0$ indicates that y_j is missing. Throughout this chapter, it is assumed that X is always fully observed. Let $F(.; \beta, X_j)$ be the CDF of $[Y_j|X_j]$, $G(.; \theta, X_j)$ be the CDF of $[R_j|X_j]$ and $\Omega(\gamma, X)$ be a correlation matrix. Particularly, $G(1; \theta, X_j) = 1$ and $G(0; \theta, X_j) = \Pr[R_j = 0|\theta, X_j]$. Then the complete CDF based on a multivariate normal copula is given by

$$\begin{aligned} H_c(y, r|X) &= \\ &\quad \Phi[\Phi^{-1}(F(y_1)), \ldots, \Phi^{-1}(F(y_m)), \Phi^{-1}(G(r_1)), \ldots, \Phi^{-1}(G(r_m))|0, \Omega] \\ &= \Phi[\Phi^{-1}(F(y)), \Phi^{-1}(G(r))|0, \Omega], \end{aligned} \tag{3.1}$$

where $\Phi^{-1}(F(y)) = (\Phi^{-1}(F(y_1)), \ldots, \Phi^{-1}(F(y_m)))$ and $\Phi^{-1}(G(r))$ is similarly defined. Now suppose that Ω can be partitioned as

$$\begin{pmatrix} \Omega_{yy} & \Omega_{yr} \\ \Omega_{ry} & \Omega_{rr} \end{pmatrix},$$

where Ω_{yy}, Ω_{rr} and Ω_{yr} are measures of the correlation among Y_i's, R_i's and that between Y_i's and R_i's, respectively. Moreover, let $u_j = \Phi^{-1}(F(y_j))$ and $u = (u_1, u_2, \ldots, u_m)^T$. Then we have

$$\begin{aligned}
\Pr[Y \le y, R = r] &= \sum_{w \le r} (-1)^{[s(r)-s(w)]} \Pr[Y \le y, R \le w] = \sum_{w \le r} (-1)^{[s(r)-s(w)]} H_c(y, w|X) \\
&= \sum_{w \le r} (-1)^{[s(r)-s(w)]} \Phi(\Phi^{-1}(F(y)), \Phi^{-1}(G(w))|0, \Omega) \\
&= \sum_{w \le r} (-1)^{[s(r)-s(w)]} \int_{-\infty}^{u} [\phi(\mu_y|0, \Omega_{yy}) \int_{-\infty}^{\Phi^{-1}(G(w))} \phi(\mu_r|\lambda_y, \Sigma) d\mu_r] d\mu_y,
\end{aligned}$$

where $\lambda_y = \Omega_{ry}\Omega_{yy}^{-1}\mu_y$, $\Sigma = \Omega_{rr} - \Omega_{ry}\Omega_{yy}^{-1}\Omega_{yr}$ and $\phi(.|\mu, \Delta)$ is the multivariate normal density function with mean μ and covariance matrix Δ. The function $s(x)$ takes the sum of the elements in x. The sum in the equation above ranges over all vectors w of length m whose elements are either 0 or 1, such that $w_j \le r_j$ for all $j = 1$ to m. Now, let $D_j = u_j'(y_j)$ and $\lambda = \Omega_{ry}\Omega_{yy}^{-1}u$. Then the complete likelihood function can be written as

$$\begin{aligned}
L_c(\beta, \theta, \gamma) &= \frac{\partial^m}{\partial y_1 \ldots \partial y_m} \Pr[Y \le y, R = r] \\
&= [\prod_{j=1}^{m} \frac{\partial u_j}{\partial y_j}] \phi(u|0, \Omega_{yy}) \sum_{w \le r} (-1)^{[s(r)-s(w)]} \int_{-\infty}^{\Phi^{-1}(G(w))} \phi(\mu_r|\lambda, \Sigma) d\mu_r \\
&= [\prod_{j=1}^{m} D_j][\phi(u|0, \Omega_{yy})][\sum_{w \le r} (-1)^{[s(r)-s(w)]} \Phi(\Phi^{-1}(G(w))|\lambda, \Sigma)] \qquad (3.2) \\
&= D(\beta)Q(\beta, \gamma)V(\beta, \theta, \gamma). \qquad (3.3)
\end{aligned}$$

It can be shown that DQ represents the distribution of $[Y|X]$ and V plays the role of selection, that is, $[R|X, Y]$.

If we let y_o be the observed values of length k with the jth element $y_{o,j}$ and Ω_o be the submatrix in Ω that corresponds to the observed values, including r, then the observed CDF is of the form:

$$H_o(y_o, r|X) = \Phi[\Phi^{-1}(F(y_o)), \Phi^{-1}(G(r))|0, \Omega_o]. \qquad (3.4)$$

We can then define $\Omega_{o,yy}$, $\Omega_{o,yr}$, $\Omega_{o,ry}$, $\Omega_{o,rr}$, $u_{o,j}$, $D_{o,j}$, λ_o and Σ_o as in the complete likelihood function. Then we have

$$L_o(\beta,\theta,\gamma) = [\prod_{j=1}^{k} D_{o,j}][\phi(u_o|0,\Omega_{o,yy})][\sum_{w\leq r}(-1)^{[s(r)-s(w)]}\Phi(\Phi^{-1}(G(w))|\lambda_o,\Sigma_o)] \tag{3.5}$$

$$= D_o(\beta)Q_o(\beta,\gamma)V_o(\beta,\theta,\gamma). \tag{3.6}$$

Therefore the total observed likelihood is

$$L_{total}(\beta,\theta,\gamma) = \prod_{i=1}^{n} L_o^{(i)} = \prod_{i=1}^{n} D_o^{(i)}(\beta)Q_o^{(i)}(\beta,\gamma)V_o^{(i)}(\beta,\theta,\gamma), \tag{3.7}$$

where $i = 1,\ldots,n$ is the subject index.

Note that in equation (3.6), V_o provides the extra term used to adjust the likelihood function in order to obtain an unbiased estimator. It is known that a wrongly specified missing-data process can also cause a serious bias problem. In Section 3.4, a simulation study is undertaken to show how this model performs under different missing-data mechanisms.

3.3.2 ESTIMATION

Meester & Mackey (1994) applied a standard maximum likelihood estimation procedure to estimate the dependence parameters of a copula and the marginal parameters simultaneously. In this chapter a pseudo maximum likelihood method is used to estimate the parameters in (3.7) (Gong & Samaniego, 1981; Parke, 1986). The basic idea of Gong & Samaniego's approach is that, to estimate a parameter of interest in the presence of a nuisance parameter, one can replace the nuisance parameter with a consistent estimator and maximize the likelihood as a function of the parameter of interest. The resulting estimator is consistent and asymptotically normal with asymptotic variance-covariance matrix properly adjusted based on the variance-covariance matrix of the consistent estimator of the nuisance parameter.

To be more specific, let the likelihood function $L_n(\delta,\pi)$ for a sample of size n be defined over two parameter vectors, δ and π with true values δ_0 and π_0, respectively. Here π is a nuisance parameter. The information matrix I for the vector $(\delta,\pi)^T$ can be partitioned as

$$\begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}.$$

Suppose that $\tilde{\pi}_n$ is a consistent estimator of π such that

$$\sqrt{n}(\tilde{\pi}_n - \pi_0) \xrightarrow{d} N(0, \Sigma).$$

Then the estimator $\hat{\delta}_n$ of δ that maximizes $L_n(\delta, \tilde{\pi}_n)$ is consistent and asymptotically normal:

$$\sqrt{n}(\hat{\delta}_n - \delta_0) \xrightarrow{d} N(0, I_{11}^{-1} + I_{11}^{-1} I_{12} \Sigma I_{21} I_{11}^{-1}). \tag{3.8}$$

In our case, we can treat θ in (3.7) as a nuisance parameter and replace it with a consistent estimator $\tilde{\theta}$. Recall that θ is associated with the marginal distribution of R_j given X_j. Since it is assumed that X_j is fully observed, the GEE method (Liang & Zeger, 1986) is used to obtain $\tilde{\theta}$ with R_j as the binary outcome and X_j as the covariate vector. Then we can replace θ with $\tilde{\theta}$ in (3.7) and maximize the corresponding likelihood function to obtain $(\hat{\beta}, \hat{\gamma})$. In this case, the Σ is the sandwich type variance-covariance matrix from the GEE and I_{11} and I_{12} can be calculated by taking the second derivative of the negative log-likelihood function with respect to β and γ. The asymptotic variance-covariance matrix of $(\hat{\beta}, \hat{\gamma})$ is then calculated based on equation (3.8).

For real data in Section 3.5 the Nelder-Mead simplex method (Nelder & Mead, 1965) is used to compute $\hat{\beta}$ and $\hat{\gamma}$. The major computational burden is the evaluation of V_o since there are many permutations to carry out when the number of repeated measures is large. Numerical integration can be used to evaluate the CDF of the multivariate normal distributions. An alternative is the Monte Carlo method, which is feasible since sampling from a multivariate normal distribution has been standardized in many software packages. To choose the starting point for the Nelder-Mead optimization, the estimates from the standard maximum likelihood procedure ignoring the missingness provide a reasonable choice. For inference, numerical differentiation is used to obtain the information matrix since the hessian matrix is not automatically computed when using the Nelder-Mead method.

For the simulation study described in next section, Newton-Raphson and trust region based algorithms (Gay, 1983) are used to maximize the likelihood function. The Newton-Raphson algorithm is first carried out. If it fails to converge, then the trust region algorithm is used.

3.4 A SIMULATION STUDY

A simulation study is conducted to evaluate the performance of the copula model when the true missing-data process deviates from that specified by the copula model. Consider a simple linear regression:

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where ϵ follows a normal distribution with mean 0 and variance σ^2. Suppose y is subject to missingness and x is always observed. If we let γ be the off diagonal element of the correlation matrix Ω, then the missing-data process defined by a bivariate normal copula function is

$$\Pr[R = 1|x, y] = \Phi\left[\frac{-\Phi^{-1}(1 - \Pr[R = 1|x]) + \gamma u}{\sqrt{1 - \gamma^2}}\right], \tag{3.9}$$

where $u = (y - \beta_0 - \beta_1 x)/\sigma$. Note that $\Pr[R = 1|x]$ can be any function of x whose range falls in [0,1].

The following three missing-data processes are considered:

$$\Pr[R = 1|x] = \text{logit}^{-1}(\theta_0 + \theta_1 x) \text{ and } (3.9) \tag{3.10}$$

$$\Pr[R = 1|x] = \arctan(\theta_0 + \theta_1 x)/\pi + 0.5 \text{ and } (3.9) \tag{3.11}$$

and

$$\Pr[R = 1|x, y] = 0.5I(l \leq 0)e^l + I(l > 0)(1 - 0.5e^{-l}), \tag{3.12}$$

where $l = \theta_0 x + \theta_1 y$. We will refer to (3.10) as the copula model with a logistic missing-data process (COPLOGI), (3.11) as the copula model with an atan missing-data process (COPATAN) and (3.12) as the exponential missing-data process (EXPONEN). Throughout the simulation study, the covariate X is generated from a standard normal distribution.

In the study we compare four different estimation procedures. They will be denoted as follows: (i) the complete cases analysis (CC) which consists of simple linear regression on the observed data; (ii) a copula model with pseudo maximum likelihood estimation (CPPML) which assumes model (3.10), estimating the θ's first and then plugging the estimates back into the likelihood function to estimate other parameters; (iii) a copula model with maximum

likelihood estimation (CPML) which assumes model (3.10) and simultaneously estimates all parameters and (iv) true maximum likelihood estimation (TRML), which correctly specifies the missing-data process (assuming θ's are known). Therefore, for model (3.10), the CPPML and CPML correctly specify the functional form of the missing-data process; for model (3.11) they correctly specify the missing-data process in the sense of (3.9), but not $[R|X]$; and for model (3.12) the CPPML and CPML entirely misspecify the missing-data process.

In the simulation study, $\beta_0 = \beta_1 = \sigma^2 = 1$ and $\gamma = 0.5$. The values for the θ's are chosen so that for each of the three missing-data processes approximately 65% of the y_i's will be observed. The simulation results are presented in Table 3.1 and 3.2, which are based on a sample size of 500 and 1000, respectively (1000 replicates). It can be seen that under all three missing-data processes the CC is seriously biased. Specifically, β_0 is always overestimated and σ^2 is always underestimated. The CPPML and CPML perform quite similarly in general and occasionally the CPPML has slightly smaller MSE than the CPML. For model (3.10), the CPPML, CPML and TRML perform equally well since all of them correctly specify the missing-data process. For model (3.11) the CPPML and CPML have slightly higher standard errors when compared with the TRML, which is due to the wrongly specified margin $[R|X]$. But on average they are still close to the correct values. The CPPML and CPML still perform well even in (3.12), when the missing-data mechanism belongs to a totally different class. As the sample size increases, the bias of the CPPML, the CPML and the TRML tends to be smaller, except for the COPATAN process. Moreover, the coverage probabilities of the CPML and CPPML are closer to 95%, though the rate of convergence seems to be slow. In summary, the CPPML and CPML are satisfactory in reducing bias. The loss of efficiency as compared to the TRML is due to the deviation of the assumed missing-data process from the true one.

Essentially (3.9) states that the selection process defined by a bivariate normal copula is of the form:

$$\Pr[R = 1|x, y] = \Phi(ay + f(x)). \tag{3.13}$$

Thus the probability of being observed is a monotone function of the outcome y and depends on x and y through the sum of an arbitrary function of x and a linear function of y. As a

matter of fact, the EXPONEN process has a shape similar to (3.13), which is why the CPPML and CPML have ignorable bias. Although most missing-data processes are complicated, we often have some "qualitative" clue as to how the values of the outcomes affect the probability of their being observed. For example, it might be reasonable to assume that subjects who perform poorly in a cognitive test tend to be more likely to refuse to take the test. Therefore (3.9) provides a class of missing-data processes for consideration when such a "qualitative" assumption can be made.

3.5 APPLICATION TO THE VERBAL FLUENCY TEST (VETFA)

We apply the proposed method to a dataset from the Monongahela Valley Independent Elder Survey (MoVIES). MoVIES is a prospective epidemiologic study of dementia, investigating incidence, risk factors and outcomes of late-life dementia, including Alzheimer's disease. The study started in 1987 and ended in 2002. The study cohort of 1681 subjects from southwestern Pennsylvania was reassessed on average every 2 years in a series of data collection waves. Attrition between waves was due to death (on average 9%-14% between waves), dropout and relocation (on average, 2.7% between waves). In addition, some subjects skipped certain waves, resulting in non-monotone missing cases.

The response variable in our example is a psychiatric test score called Verbal Fluency: Fruits and Animals or VETFA (Lezak, 1995). This test measures impairment in verbal fluency, semantic memory and language. Each subject is asked to name things that belong to each category of fruits and animals as fast as possible within one minute. The subject's score is the number of words given in each 15-second-interval. Repetitions, improper nouns and different forms of the same instance (e.g. bear, black bear) are not counted.

Data from wave 1, wave 3 and wave 5 are included for this analysis. Thus each subject has at most 3 measurements on VETFA (VETFA has to be observed at wave 1 to be included in our analysis). The following covariates are included in the analysis: age at baseline (*age*), sex (*female*: 1-female, 0-male), education level (*highedu*: 1-high school or higher, 0-otherwise) and time (t) from baseline in years. Table 3.3 shows the distribution of the missing-data patterns. It is clear from Table 3.3 that males and subjects with less than

a high school education level tend to have more missing VETFA scores as compared with females and subjects with a higher education level.

Our experience suggests that subjects tend to drop out of the study or refuse to take the test as their test scores get lower. Or in other words, subjects who have all three test scores available are relatively healthier. Therefore the missing-data process is assumed to be non-ignorable. To connect the mean of the response and the covariates, an additive model and a multiplicative model were fit to the data.

3.5.1 ADDITIVE MODEL

A linear mixed effects model is combined with the copula structure. Let Z be the submatrix of X that contains the intercept and time from baseline. Then a mixed effects model states

$$Y = X\beta + Z\alpha + \epsilon, \tag{3.14}$$

where $\alpha = (\alpha_0, \alpha_t)^T \sim N(0, \Psi)$ and $\epsilon \sim N(0, \sigma^2 I)$. Thus Ω_{yy} is determined by the model above. To model Ω_{yr} and Ω_{rr}, the following correlation structure was used:

$$\omega_{yr}^{ij} = \exp(-\eta(\delta + |t_i - t_j|)) \tag{3.15}$$

$$\omega_{rr}^{kj} = \exp(-\tau|t_k - t_j|), \tag{3.16}$$

where $i = 1, 2$ or 3 and $j, k = 2$ or 3 since response is always observed at baseline. Thus in the framework of (3.7), $(\beta, \gamma) = (\beta, \eta, \delta, \tau, \Psi, \sigma^2)$. To estimate these parameters, GEE was first applied to $R = (R_2, R_3)^T$ with baseline age, sex, education level and time from baseline as the covariates. The resulting estimates were then "plugged" into the copula-based likelihood function. Parameter estimates from the GEE (not shown) confirm again that males and subjects with a lower education level tend to have a greater number of missing outcomes. Moreover, subjects who are older at baseline are more likely to have missing outcomes as expected.

For the response model, a main effects model and a second model with two interaction terms $age*t$ and $female*t$ (the only two two-way interaction terms that are significant from a mixed effects model ignoring the missingness) were fit to the data. Results from the mixed

effects model and the normal copula-based selection model are shown in Table 3.4. It is clear that the major difference between the mixed effects model and the copula selection model is the slope of time. Ignoring the missingness results in a slower decline rate. Intuitively, since subjects with lower outcomes tend to drop out of the study, regression on the observed data tend to produce a less steep slope for time. Similarly, since male subjects and subjects who are older at baseline tend to drop from the study due to lower scores, the absolute values of $\beta_{female*t}$ and β_{age*t} from the mixed effects model should be underestimated as is the case in Table 3.4. Moreover, η, δ and τ provide a measure of the correlation between the outcomes and the missing-data indicators. For example, $\omega_{yr}^{ii} = \exp(-0.032 \times 36.78) = 0.31$ based on the main effects model.

3.5.2 MULTIPLICATIVE MODEL

One of the major advantages of copula based parametric models is that one can fit any marginal distribution without changing the correlation structure. In a second experiment, a gamma model for the marginal distribution was fit to the data. Specifically,

$$[Y_j|X_j] \sim \text{Gamma}(\mu_j, \nu_j) \tag{3.17}$$

$$\log\mu_j = X_j\beta. \tag{3.18}$$

Here μ_j is the mean and ν_j is the scale parameter for $j = 1, 2$ or 3. Therefore, each of the three time points has its own scale parameter. It is in the sense of the link (3.18) that we call the model multiplicative. To model Ω_{yy}, consider the same model as (3.16):

$$\omega_{yy}^{ij} = \exp(-\rho|t_i - t_j|) \tag{3.19}$$

for $i, j = 1, 2$ or 3. In this analysis, Ω_{yr} and Ω_{rr} are the same as in the additive model. Then in the framework of (3.7), $(\beta, \gamma) = (\beta, \nu_1, \nu_2, \nu_3, \rho, \eta, \delta, \tau)$. The results are shown in Table 3.5. Again, we see a similar pattern as for the additive model in β_t, β_{age*t} and $\beta_{female*t}$. Thus this example clearly demonstrates how the bias can be reduced by a copula selection model.

In summary, females and younger subjects tend to have higher VETFA scores. Moreover, the performance of older subjects tend to decline faster. Ignoring the missing-data process in this example results in underestimating the decline rate of the score.

3.6 DISCUSSION

In this chapter, a normal copula based selection model is proposed for continuous responses subject to non-ignorable non-monotone missing-data processes. A simulation study is carried out to compare the performance of the model under different missing-data processes and the method is applied to a real dataset. Essentially, the normal copula specifies a particular class of selection models defined by V in (3.3). The concept is straight forward and the method is easy to implement in practical application. The model is related to Heckman's (1976) probit selection model and the Tobit model by Amemiya (1984). From the likelihood function (3.5 and 3.6) it can be seen that if $\Omega_{yr} = 0$ then the missing-data mechanism is ignorable. In our example, this can be checked by testing that $\eta\delta$ is larger than some big number.

The major computational burden of the proposed method is the calculation of V_o. Essentially, the normal copula-based selection model reduces the integration of "missing outcomes" for selection models to the evaluation of the CDF of a multivariate normal distribution at different points. When the number of repeated measurements is large, evaluation of the likelihood function may take a substantial amount of time. Since many software packages have subroutines for sampling from the multivariate normal distribution, Monte Carlo methods provide an alternative method to numerical integration.

The "Shared parameter" selection model has been a popular tool to model non-ignorable non-monotone missing outcomes. However, it is often hard to see directly how the outcomes and the missing-data indicators are correlated since they are connected by a random parameter. Copula based selection models provide a more clear insight on how the outcome drives the missing-data process. Moreover, it allows for flexibility in the choice of both the margins and the copula that defines the missing-data process.

Table 3.1: Simulation results (sample size=500, 1000 replicates) for the complete case analysis (CC), copula model with (pseudo) maximum likelihood estimation (CPPML and CPML) and MLE by correctly specifying the missing-data process (TRML) under COPLOGI (see (3.10)), COPATAN (see (3.11)) and EXPONEN (see (3.12))

bias×100 S.E.×100 $\sqrt{\text{MSE}}$×100 95% cover prb	COPLOGI				COPATAN				EXPONEN			
	CC	CPPML	CPML	TRML	CC	CPPML	CPML	TRML	CC	CPPML	CPML	TRML
β_0	26.9	1.6	1.4	1.4	27.8	-0.2	-0.2	1.6	32.4	2.8	2.8	-0.2
	5.9	15.9	15.9	15.7	5.8	17.5	17.5	16.6	5.5	14.8	15.4	5.9
	27.5	16.0	16.0	15.8	28.4	17.5	17.5	16.6	32.8	15.1	15.6	5.9
	0.005	0.913	0.913	0.905	0.002	0.896	0.895	0.901	0.000	0.920	0.912	0.954
β_1	19.3	1.2	1.1	1.1	-16.9	0.8	0.9	-1.0	-24.3	-2.7	-2.7	0.2
	6.4	12.0	12.0	12.0	6.5	12.1	12.2	11.5	6.4	11.9	12.3	6.3
	20.3	12.1	12.1	12.1	18.1	12.2	12.2	11.5	25.1	12.2	12.6	6.3
	0.157	0.930	0.928	0.926	0.253	0.916	0.914	0.919	0.046	0.932	0.929	0.947
σ^2	-8.9	0.5	0.6	0.6	-10.1	2.1	2.2	0.8	-12.4	-0.7	-0.6	-0.7
	7.0	10.9	10.9	10.8	7.1	12.6	12.6	12.2	6.8	11.2	11.3	8.3
	11.3	10.9	11.0	10.9	12.3	12.8	12.8	12.2	14.2	11.2	11.3	8.4
	0.736	0.938	0.938	0.935	0.692	0.935	0.933	0.914	0.557	0.920	0.914	0.943

Table 3.2: Simulation results (sample size=1000, 1000 replicates) for the complete case analysis (CC), copula model with (pseudo) maximum likelihood estimation (CPPML and CPML) and MLE by correctly specifying the missing-data process (TRML) under COPLOGI (see (3.10)), COPATAN (see (3.11)) and EXPONEN (see (3.12))

bias×100 S.E.×100 $\sqrt{\text{MSE}}$×100 95% cover prb	COPLOGI				COPATAN				EXPONEN			
	CC	CPPML	CPML	TRML	CC	CPPML	CPML	TRML	CC	CPPML	CPML	TRML
β_0	27.1	1.1	1.0	1.1	28.2	-1.2	-1.3	0.8	32.8	1.5	1.2	0.0
	4.0	11.0	10.9	10.9	4.0	12.5	12.4	11.2	4.0	9.2	9.1	4.2
	27.4	11.1	10.9	11.0	28.5	12.6	12.5	11.2	33.0	9.3	9.2	4.2
	0.000	0.933	0.935	0.934	0.000	0.916	0.914	0.928	0.000	0.955	0.952	0.953
β_1	19.3	0.7	0.7	0.7	-17.2	1.4	1.6	-0.5	-24.6	-1.7	-1.5	0.0
	4.6	8.8	8.7	8.8	4.6	9.0	8.9	8.0	4.6	7.8	7.7	4.5
	19.8	8.8	8.7	8.8	17.8	9.1	9.0	8.0	25.0	8.0	7.8	4.5
	0.016	0.931	0.931	0.924	0.036	0.922	0.920	0.940	0.000	0.946	0.945	0.949
σ^2	-8.6	0.2	0.2	0.2	-9.8	2.1	2.1	0.5	-12.4	-0.6	-0.4	-0.3
	4.9	7.9	8.0	7.8	5.1	9.2	9.3	9.0	4.8	7.9	8.0	6.0
	9.9	7.9	8.0	7.8	11.0	9.4	9.5	9.0	13.3	7.9	8.0	6.0
	0.602	0.935	0.933	0.939	0.508	0.935	0.934	0.919	0.298	0.930	0.929	0.945

Table 3.3: Distribution of the missing-data patterns of VETFA

pattern	baseline	wave 3	wave 5	male(%)	female(%)	lowedu(%)	highedu(%)	total(%)
1	•	•	•	251(35.70)	485(50.31)	258(35.98)	478(50.32)	736(44.2)
2	•	•	×	159(22.62)	218(22.61)	169(23.57)	208(21.89)	377(22.6)
3	•	×	•	11(1.56)	18(1.87)	17(2.37)	12(1.26)	29(1.7)
4	•	×	×	282(40.11)	243(25.21)	273(38.08)	252(26.53)	525(31.5)
total				703(100)	964(100)	717(100)	950(100)	1667(100)
•: observed, ×: missing								

Table 3.4: Parameter estimates from the mixed effects model and the copula selection model under the additive mean structure

parameters	mixed effects model						copula selection model					
	main effects			with interaction			main effects			with interaction		
	est	S.E.	p-value	est	S.E.	p-value	est	S.E.	p-value	est	S.E.	p-value
β_0	24.18	0.26	< 0.0001	24.27	0.26	< 0.0001	24.13	0.35	< 0.0001	24.28	0.29	< 0.0001
β_{age}	-0.42	0.02	< 0.0001	-0.38	0.02	< 0.0001	-0.41	0.03	< 0.0001	-0.38	0.03	< 0.0001
β_{female}	0.62	0.28	0.0292	0.54	0.29	0.0643	0.64	0.31	0.0390	0.54	0.31	0.0815
β_{edu}	2.53	0.29	< 0.0001	2.51	0.29	< 0.0001	2.45	0.40	< 0.0001	2.52	0.34	< 0.0001
β_t	-0.36	0.02	< 0.0001	-0.46	0.04	< 0.0001	-0.46	0.02	< 0.0001	-0.55	0.04	< 0.0001
β_{age*t}				-0.032	0.005	< 0.0001				-0.038	0.005	< 0.0001
$\beta_{female*t}$				0.070	0.045	0.118				0.082	0.048	0.109
η							0.032	0.008		0.030	0.006	
δ							36.78	9.12		40.18	9.01	
τ							0.035	0.006		0.031	0.005	
-2ll	22011.6(REML)			21975.4(REML)			19048.1			18984.0		

Table 3.5: Parameter estimates from the gamma copula model and the copula selection model under the multiplicative mean structure

parameters	gamma copula model						copula selection model					
	main effects			with interaction			main effects			with interaction		
	est	S.E.	p-value	est	S.E.	p-value	est	S.E.	p-value	est	S.E.	p-value
β_0	3.20	0.012	< 0.0001	3.19	0.012	< 0.0001	3.18	0.029	< 0.0001	3.19	0.095	< 0.0001
β_{age}	-0.018	0.001	< 0.0001	-0.016	0.001	< 0.0001	-0.019	0.002	< 0.0001	-0.016	0.004	< 0.0001
β_{female}	0.014	0.012	0.243	0.018	0.012	0.134	0.028	0.038	0.461	0.019	0.188	0.919
β_{edu}	0.089	0.013	< 0.0001	0.082	0.012	< 0.0001	0.094	0.042	0.0252	0.085	0.150	0.571
β_t	-0.014	0.001	< 0.0001	-0.018	0.002	< 0.0001	-0.019	0.004	< 0.0001	-0.023	0.003	< 0.0001
β_{age*t}				-0.0018	0.0002	< 0.0001				-0.0022	0.0008	0.006
$\beta_{female*t}$				0.00020	0.0022	0.920				0.00051	0.0198	0.980
ρ	0.075	0.004		0.079	0.003		0.084	0.001		0.076	0.068	
η							0.048	0.005		0.039	0.012	
δ							25.43	0.36		29.21	13.52	
τ							0.044	0.002		0.048	0.02	
-2ll	15952.0			15866.8			19509.9			19414.0		

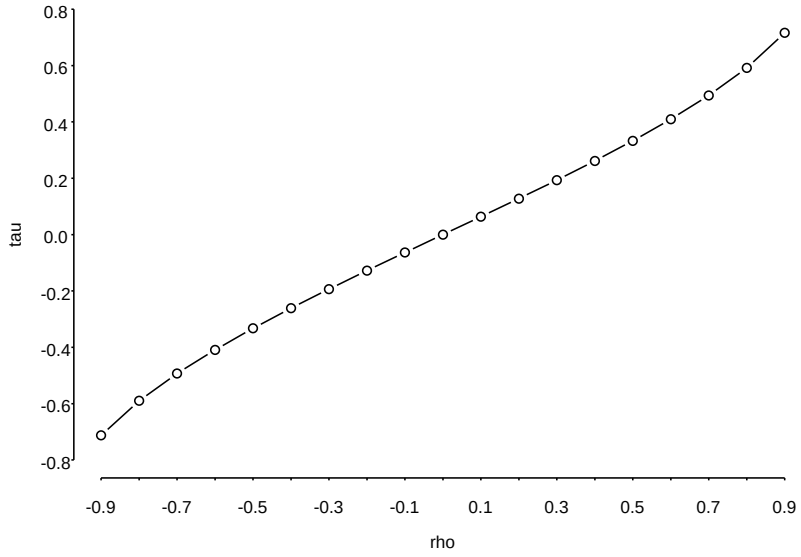

Figure 3.1: rho vs. tau for the bivariate normal copula

4.0 PATTERN-MIXTURE MODEL WITH PSEUDO MAXIMUM LIKELIHOOD ESTIMATION

In this chapter, the focus is on the development of pattern-mixture models for data with a non-ignorable missing response. In estimating the regression parameters that are identifiable, we use the pseudo maximum likelihood method introduced by Gourieroux et al. (1984). This procedure provides consistent estimators when the mean structure is correctly specified for each pattern, with further information on the variance structure giving an efficient estimator. A "three step" estimation approach is used to calculate the efficient estimator. The proposed method can be used to handle a variety of continuous and discrete outcomes. A Hausman type test (Hausman, 1978) of model misspecification is also developed for model simplification in order to improve efficiency. Throughout this chapter, it is assumed that models are already identifiable and the focus is on parameter estimation and simplification of such models. It should be emphasized that this approach by no means indicates a complete strategy for the fitting of pattern-mixture models. The idea is to provide an extension to the current pattern-mixture model approach in terms of estimation and efficiency improvement. Although the focus is monotone missing data in this chapter, the approach can be applied to data with general missing-data patterns directly. Pseudo maximum likelihood estimation and the test of model misspecification are introduced in Sections 4.1.1 and 4.1.2, respectively. Simulation studies are presented in Section 4.2 to evaluate the performance of the proposed estimation method and the power of the proposed test. The proposed method is applied to an epidemiologic cohort study to examine cognition decline among elderly in Section 4.3. Finally, we conclude this chapter with a discussion section.

4.1 APPLICATION OF PSEUDO MAXIMUM LIKELIHOOD ESTIMATION TO PATTERN-MIXTURE MODELS

4.1.1 PSEUDO MAXIMUM LIKELIHOOD

One of the difficulties encountered in fitting pattern-mixture models is that the distribution for each missing-data pattern can be very hard to model parametrically. This is due to the complexity of the missing-data process. A simple example is that the conditional distribution for a certain missing-data pattern is not normal even we can reasonably assume that the unconditional distribution is normal, unless the missing-data process depends on the outcome in a specific manner. Therefore, a semi-parametric pattern-mixture model that does not require the modelling of the exact distributional form is desirable. In this subsection, a pseudo maximum likelihood based estimation procedure is proposed within the framework of pattern-mixture models requiring only the specification of the first (and second) moment(s) for each pattern.

Gourieroux et al. (1984) proposed a regression model with specification of the mean structure up to unknown parameters. They then assumed the conditional distribution of the response given the covariates follows some distribution that belongs to an exponential family. They proved that the corresponding Pseudo Maximum Likelihood Estimator (PMLE) that maximizes the assumed likelihood function is consistent even though the true distribution does not belong to the proposed family, provided that the mean structure is correctly specified. Moreover, some exponential families have an additional nuisance parameter $\eta = g(\mu, \Sigma)$, where μ and Σ are the mean and covariance, respectively. Here g defines a one to one function of Σ for any fixed μ. The normal distribution with mean μ and variance σ^2, is a typical example of this type of distribution with $\eta = \sigma^2$. Thus if we know the true functional form of the variance and there are consistent estimators of the parameters associated with the mean and variance, then we can consistently estimate η. Gourieroux et al. (1984) showed that the estimator based on an exponential family with nuisance parameters replaced by corresponding consistent estimators is efficient over the PMLE. They named this estimator the Quasi Generalized Pseudo Maximum Likelihood Estimator (QGPMLE). Gourieroux et

al.'s approach can be seen as a multi-dimensional version of the quasi-likelihood functions proposed by Wedderburn (1974) and is closely connected to the GEE (Liang & Zeger, 1986).

To illustrate the idea, consider a simple regression problem $E(Y|X) = \theta X$, where $X, \theta > 0$, and Y takes non-negative integer values. Then we can assume a normal distribution $p(Y|x) = \phi(y; x, \theta) = N(\theta x, 1)$ of Y given X, which results in the same estimator as the least square estimator, $\hat{\theta}_{NOR} = \sum x_i y_i / \sum x_i^2$. Alternatively we can also assume a Poisson distribution, $p(Y|x) = \lambda(y; x, \theta) = POI(\theta x)$, which results in $\hat{\theta}_{POI} = \bar{y}/\bar{x}$. If the proposed mean structure is correct with true value θ_0, then both $\hat{\theta}_{NOR}$ and $\hat{\theta}_{POI}$ will converge to θ_0 regardless of the true distribution of $[Y|X]$ since the normal distribution with known variance and the Poisson distribution are exponential families.

Now suppose we also know that $Var(Y|x) = \alpha h(x)$ with h known. Then α can be consistently estimated as

$$\tilde{\alpha} = \sum_{i=1}^{n} \frac{(y_i - x_i \hat{\theta}_{NOR})^2}{h(x_i)} / n.$$

Then we assume that the variance of $[Y_i|x_i]$ is $\tilde{\alpha} h(x_i)$ and again assume a normal distribution to obtain the efficient estimator $\hat{\theta}_{QGPMLE}$ by maximizing the corresponding pseudo likelihood function. Therefore a strategy to compute the QGPMLE is composed of three steps (Lipsitz et al., 1992): (i) obtain a PMLE, (ii) estimate the nuisance parameter based on the PMLE and (iii) calculate the QGPMLE. Step (ii) is not trivial in general. However, under some model structure it is indeed quite straight forward.

To apply the above estimation procedure to pattern-mixture models, consider a single observation with $y = (y_1, y_2, \ldots, y_m)^T$ an $m \times 1$ outcome vector and X an $m \times p$ covariate matrix. We temporarily drop the subject index for the sake of clarity. Suppose y is subject to a monotone missing-data process and let R be the missing-data pattern indicator, where $R = r$ indicates that the subject has r observed outcomes. Let $y_{(j)}$ and $X_{(j)}$ be the vector composed of the first j elements of y and the matrix composed of the first j rows of X, respectively. It is assumed that $X_{(r)}$ is always fully observed. In a regression analysis, the primary interest is $E[Y|X, \theta, \phi]$, which can be written as

$$E[Y|X, \theta, \phi] = \sum_{r=1}^{m} E[Y|X, r, \theta] p(r|X, \phi), \tag{4.1}$$

where $p(r|X,\phi)$ represents the probability mass function for $[R|X]$. Here it is assumed that θ and ϕ are distinct. As mentioned at the beginning of this chapter, it is also assumed that θ is identifiable. Note by the way equation (4.1) is presented, we avoid specification of the distributional form of Y due to the semi-parametric nature of the approach that is going to be proposed.

For simplicity, it is assumed that the distribution of R conditional on X only depends on time-independent covariates, which is fully observed. Then multinomial regression can be used to estimate ϕ if the number of observed missing-data patterns is not great. For θ, inference is based on the specification of the first two moments of the conditional distribution of $Y_{(r)}$:

$$E[Y_{(r)}|X,r,\theta] = f_r(X_{(r)},\theta(r)), \tag{4.2}$$

$$Var[Y_{(r)}|X,r,\alpha] = \Omega_r(X_{(r)},\alpha(r)). \tag{4.3}$$

Here $\theta = \bigcup_r \theta(r) \in \Theta \subset R^d$ and $\alpha = \bigcup_r \alpha(r) \in \Lambda \subset R^q$, where $\theta(r)$ and $\alpha(r)$ are the subsets of θ and α that are associated with the mean and variance specification of pattern r. For fixed r, $\theta(r)$ and $\alpha(r)$ need not be distinct. Furthermore, $\theta(r)$'s and $\alpha(r)$'s need not be distinct across patterns. Note it is assumed that the distribution of $Y_{(r)}$ does not depend on future X values conditional on $X_{(r)}$.

Now suppose there are n_r subjects from pattern r and let $n = \sum_{r=1}^{m} n_r$. Let y_{rj} and X_{rj} be the observed outcome vector and corresponding covariate matrix for subject j in pattern r. Then the following two theorems are a direct extension of Theorems 3 and 4 in Gourieroux et al. (1984). The proof is given in the appendix.

THEOREM 1. Consistency and Asymptotic Normality of PMLE
Let $e_r(y,\mu_r) = \exp(A_r(\mu_r) + B_r(y) + C_r(\mu_r)y)$ *be an exponential family density function on* $\boldsymbol{R}^r$ *with mean* μ_r*, where* A_r*,* B_r *are scalars and* C_r *a row vector of size* r*. Then under regularity conditions the estimator* $\hat{\theta}_n$ *of* θ_0 *that maximizes*

$$\sum_{r=1}^{m}\sum_{j=1}^{n_r} \log[e_r(y_{rj}, f_r(X_{rj},\theta(r))]$$

is consistent and asymptotically normal:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, J^{-1}IJ^{-1}),$$

$$\begin{aligned} J &= E\left(\frac{\partial f_R}{\partial \theta_0}(\Sigma_R^{\theta_0})^{-1}\frac{\partial f_R}{\partial \theta_0'}\right), \\ I &= E\left(\frac{\partial f_R}{\partial \theta_0}(\Sigma_R^{\theta_0})^{-1}\Omega_R(\Sigma_R^{\theta_0})^{-1}\frac{\partial f_R}{\partial \theta_0'}\right), \end{aligned}$$

where $\Sigma_r^{\theta_0}(X)$ *is the covariance matrix associated with* $e_r(., f_r(X_{(r)}, \theta_0(r)))$.

Theorem 2. Consistency and Asymptotic Normality of QGPMLE
Suppose the functional form of Ω_r *is known. Let* $e_r^*(y, \mu_r, \eta_r) = \exp(A_r(\mu_r, \eta_r) + B_r(\eta_r, y) + C_r(\mu_r, \eta_r)y)$ *be a density function on* $\boldsymbol{R}^r$ *with mean* μ_r *and* $\eta_r = g_r(\mu_r, \Sigma_r)$, *where* Σ_r *is the covariance matrix. Let* $\tilde{\theta}_n$ *and* $\tilde{\alpha}_n$ *be strongly consistent estimators of* θ_0 *and* α_0, *such that* $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ *and* $\sqrt{n}(\tilde{\alpha}_n - \alpha_0)$ *are bounded in probability, then under regularity conditions the estimator* $\hat{\theta}_n$ *of* θ_0 *that maximizes*

$$\sum_{r=1}^{m}\sum_{j=1}^{n_r}\log[e_r^*(y_{rj}, f_r(X_{rj}, \theta(r)), g_r(f_r(X_{rj}, \tilde{\theta}_n(r)), \Omega_r(X_{rj}, \tilde{\alpha}_n(r))))]$$

is consistent and asymptotically normal:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, V = [E\left(\frac{\partial f_R}{\partial \theta_0}(\Omega_R)^{-1}\frac{\partial f_R}{\partial \theta_0'}\right)]^{-1}).$$

Moreover, V *is the lower bound of the asymptotic variance of the estimator introduced in theorem 1.*

These two theorems provide the basis for the computation of the PMLE and QGPMLE. As noted in Theorem 2, to find the QGPMLE one first needs to provide a consistent estimator $\tilde{\theta}_n$. The PMLE provides such an estimate and this estimate can be used to obtain $\tilde{\alpha}_n$ (Gourieroux et al., 1984). This estimation approach is similar to the method of Lipsitz et al. (1992). The advantage is that it simplifies a difficult estimation problem by reducing the number of parameters, bypassing the need to simultaneously obtain $\tilde{\theta}_n$ and $\tilde{\alpha}_n$. For the QGPMLE, one does not need a $\sqrt{n}$-consistent estimator for every element of α. It is enough

to have $\sqrt{n}$-consistent estimators of those that are associated with the variance specification of the observed vector for each pattern.

Under the linear mean structure framework, the PMLE and QGPMLE are of the form:

$$(\sum_{r=1}^{m} \tilde{X}_r^T W_r^{-1} \tilde{X}_r)^{-1} (\sum_{r=1}^{m} \tilde{X}_r^T W_r^{-1} \tilde{Y}_r). \tag{4.4}$$

The asymptotic variance of the PMLE and QGPMLE are consistently estimated as follows:

$$\hat{V}_1 = n(\sum_{r=1}^{m} \tilde{X}_r^T \hat{\Sigma}_r^{-1} \tilde{X}_r)^{-1} (\sum_{r=1}^{m} \sum_{j=1}^{n_r} X_{rj}^T \hat{\Sigma}_{rj}^{-1} \hat{\epsilon}_{rj} \hat{\epsilon}_{rj}^T \hat{\Sigma}_{rj}^{-1} X_{rj}) (\sum_{r=1}^{m} \tilde{X}_r^T \hat{\Sigma}_r^{-1} \tilde{X}_r)^{-1}, \tag{4.5}$$

$$\hat{V}_2 = n(\sum_{r=1}^{m} \tilde{X}_r^T \hat{\Omega}_r^{-1} \tilde{X}_r)^{-1}. \tag{4.6}$$

Here $W_r = \hat{\Sigma}_r$ for the PMLE and $W_r = \hat{\Omega}_r$ for the QGPMLE; $\tilde{X}_r$ is the covariate matrix with respect to θ, stacking over subjects within pattern r; $\tilde{Y}_r$ is the response vector, stacking over subjects within pattern r; $\hat{\Sigma}_{rj}$ is the "working" covariance matrix; $\hat{\Sigma}_r$ is the block diagonal matrix composed of $\hat{\Sigma}_{rj}$; $\hat{\epsilon}_{rj}$ is the residual and $\hat{\Omega}_r$ is the block diagonal matrix composed of $\hat{\Omega}_{rj}$, which is the estimated true covariance matrix. We will apply equation (4.4)-(4.6) to the simulation study in Section 4.2 and the Mini Mental State Exam in Section 4.3.

For a linear pattern-mixture model, we are mainly interested in the marginal parameter estimator ($\hat{\theta}_{mar}$). This estimator can be computed as a weighted sum of the estimator of each pattern ($\hat{\theta}(r)$) with the estimated proportion of each pattern ($\hat{\pi}_r$) as the weight. Thus we have

$$\hat{\theta}_{mar} = \sum_{r=1}^{m} \hat{\pi}_r \hat{\theta}(r).$$

Here $\hat{\pi}_r$ can be calculated according to the multinomial regression model, which results in the covariates-specific marginal parameter vector; or one can use empirical proportion of pattern r as $\hat{\pi}_r$ to obtain the overall marginal parameter vector. Delta method or bootstrap approach can be used to estimate the variance of $\hat{\theta}_{mar}$.

4.1.2 MODEL MISSPECIFICATION

To limit the loss of the degrees of freedom of a model, we would want to simplify an identifiable pattern-mixture model. This is of particular significance due to the large dimension of the parameter vector. For example, if there are five patterns and for each pattern there are five parameters to be estimated, then the length of the parameter vector is 25. For such a model, inference requires each pattern to have a sufficient number of data points, which in practice might not be easy to achieve. Although a Wald type test can be used to check certain constraints on model parameters in a semi-parametric setting, a test that can be used to evaluate the overall adequacy of a model is desirable for our purpose.

Hausman (1978) proposed a test for model misspecification that can be extended to address this issue. Suppose a model is not misspecified, or in other words, there exists a $\theta_0 \in \Theta$ such that the corresponding member $f(x, \theta_0)$ of the proposed family of distributions $f(x, \theta)$ is the true distribution $g(x)$. Then under regularity conditions, the $\hat{\theta}_n$ that maximizes $\sum_i \log[f(x_i, \theta)]$ converges to θ_0. If such a θ_0 does not exist, then under regularity conditions $\hat{\theta}_n$ converges to a θ^* that minimizes the Kullback-Leibler (Kullback & Leibler, 1951) Information Criterion (KLLC):

$$I(g : f, \theta) = E_g \left\{ \log \frac{g(X)}{f(X, \theta)} \right\}.$$

For example, both $\hat{\theta}_{NOR}$ and $\hat{\theta}_{POI}$ in Section 4.1.1 will converge to the true value when the linear structure is correct. On the other hand, suppose the linear structure is not correct. Let $g(x, y)$ be the true joint distribution of (X, Y) and

$$\begin{aligned} I_{NOR} &= E_g \left\{ \log \frac{g(X, Y)}{\phi(Y; X, \theta)} \right\} \\ I_{POI} &= E_g \left\{ \log \frac{g(X, Y)}{\lambda(Y; X, \theta)} \right\}. \end{aligned}$$

Then $\hat{\theta}_{NOR}$ and $\hat{\theta}_{POI}$ will converge to θ^* and θ^{**}, respectively, where θ^* minimizes I_{NOR} and θ^{**} minimizes I_{POI}. In general, $\theta^* \neq \theta^{**}$ and we expect to detect such a difference as the sample size gets large.

Thus the general strategy is to construct two different estimators of θ that will both converge to the true value if the model is correctly specified and will converge to different limits

if the model is wrong. The PMLE and QGPMLE provide candidates for such estimators. The next theorem sets up the null distribution of a test statistic based on the PMLE and QGPMLE. For the theoretical development of this type of test, see White (1981, 1982). A sketch of the proof is shown in the appendix.

Theorem 3. Let $\hat{\theta}_n$ and $\tilde{\theta}_n$ be the estimators of theorem 1 and theorem 2 with asymptotic variance-covariance matrix V_h and V_t. Then under the assumption that the mean structure and the variance structure (for $\tilde{\theta}_n$) are correctly specified we have

$$n(\hat{\theta}_n - \tilde{\theta}_n)^T (V_h - V_t)^{-1} (\hat{\theta}_n - \tilde{\theta}_n) \xrightarrow{d} \chi^2_d,$$

provided that $V_h - V_t$ is positive definite. Moreover, the same limit distribution holds if V_h and V_t are replaced by their corresponding consistent estimators.

Thus this test provides a convenient tool to check model adequacy after the PMLE and QGPMLE are calculated. This test will be applied to an example in Section 4.3.

4.2 SIMULATION STUDIES

4.2.1 COMPARISON OF THE PMLE AND QGPMLE WITH OTHER APPROACHES

To study the properties of the proposed estimators a simulation study is conducted. The PMLE and QGPMLE are compared to estimators obtained under the MLE based on complete cases (CC), the MLE based on observed data ignoring the missing-data mechanism (MLE1), the weighted estimating equation (Rotnitzky et al., 1998) by correctly specifying the weight (WEE) and the MLE based on the correct specification of the missing-data process (MLE2). The purpose of the study is to evaluate how well the PMLE and QGPMLE do when compared to the CC and MLE1 that ignore the missing-data process and how they compare with the WEE that is another semi-parametric approach to handle missing data. Intuitively, we expect that the PMLE and QGPMLE are less biased than the CC and MLE1

under a non-ignorable missing-data process. However, efficiency is also of our interest since the PMLE and QGPMLE do not use the information of distributional forms. The difference between the WEE and PMLE (QGPMLE) is that the WEE uses information regarding the missingness in the format of the probability of being in each pattern instead of the moment specification for each pattern.

Three missing-data mechanisms considered are: (i) missing-data mechanism only depends on covariates or covariates dependent missing (CDM), (ii) missing-data process depends on observed outcome or missing at random (MAR) and (iii) missing-data process depends on unobserved outcome or missing not at random (MNAR).

To be more specific, let the random triple (X, Y_1, Y_2) be generated by the following model:

$$X \sim \mathrm{U}[0,1],\ [Y_1|x] \sim \mathrm{EXP}(\alpha x),\ [Y_2|x, y_1] \sim \mathrm{EXP}(\frac{\beta}{\alpha} y_1).$$

Here $\alpha, \beta > 0$, $\mathrm{U}[0,1]$ refers to the uniform distribution on $[0,1]$ and $\mathrm{EXP}(\mu)$ refers to the exponential distribution with mean μ. Thus $E(Y_1|x) = \alpha x$ and $E(Y_2|x) = \beta x$. Note also that given x, Y_1 and Y_2 are positively correlated.

To generate the different missing scenarios, suppose that x and y_1 are always observed and that y_2 is subject to missingness. Let R be the missing-data indicator such that $R = 1$ indicates that y_2 is observed and $R = 0$ indicates that y_2 is missing. Then the three missing-data processes are of the form:

$$\text{CDM:}\quad \Pr[R = 1|x, y_1, y_2] = x^{k_1}, \tag{4.7}$$

$$\text{MAR:}\quad \Pr[R = 1|x, y_1, y_2] = \exp(-\frac{y_1}{k_2 \alpha x}), \tag{4.8}$$

$$\text{MNAR:}\quad \Pr[R = 1|x, y_1, y_2] = \exp(-\frac{y_1}{k_3 \alpha x} - \frac{\alpha y_2}{k_3 \beta y_1}). \tag{4.9}$$

Here k_1, k_2 and k_3 are all positive. Note that although (4.8) is MAR, the missing-data process does depend on the regression parameter α. Thus the estimator of α by MLE1 is not efficient since it ignores the missing-data process. For the CDM, (Y_1, Y_2) and R are independent given X. For the MAR, observations with a higher value of y_1 have a lower probability of being complete cases, so that the mean of y_1 for complete cases should be smaller than the marginal mean and that of the incomplete cases should be larger than the

marginal mean. Since Y_1 and Y_2 are positively correlated, we should expect to see a smaller mean of Y_2 for complete cases as compared with its marginal mean. It turns out that similar observations can also be made under the MNAR as in the MAR. The mean structure for each of the two patterns is of the form:

$$\begin{aligned}
\text{CDD:}\quad E[Y_1|x,R=1] &= \alpha x \\
E[Y_2|x,R=1] &= \beta x \\
E[Y_1|x,R=0] &= \alpha x \\
\text{MAR:}\quad E[Y_1|x,R=1] &= \frac{k_2}{k_2+1}\alpha x \\
E[Y_2|x,R=1] &= \frac{k_2}{k_2+1}\beta x \\
E[Y_1|x,R=0] &= \frac{2k_2+1}{k_2+1}\alpha x \\
\text{MNAR:}\quad E[Y_1|x,R=1] &= \frac{k_3}{k_3+1}\alpha x \\
E[Y_2|x,R=1] &= \left(\frac{k_3}{k_3+1}\right)^2 \beta x \\
E[Y_1|x,R=0] &= \frac{3k_3^2+3k_3+1}{2k_3^2+3k_3+1}\alpha x.
\end{aligned}$$

For the PMLE and QGPMLE, the pseudo likelihood function is based on the (bivariate) normal distribution. For the PMLE, $\hat{\Sigma}_{rj} = \mathbf{I}_r$, where $\mathbf{I}_r$ is the r-dimensional identity matrix (see (4.4) and (4.5)). The PMLE is then used to estimate the QGPMLE (see Theorem 2). For the PMLE and QGPMLE it is assumed that k_2 and k_3 are known. In the simulation, $\alpha = 2$ and $\beta = 4$. Approximately 50% of the y_2 values will be missing by setting $k_1 = 1$, $k_2 = 1$ and $k_3 = 2.4$, and 25% of the y_2 values will be missing by setting $k_1 = 0.3$, $k_2 = 3$ and $k_3 = 6.5$.

Tables 4.1 and 4.2 show the simulation results based on 1000 replicates. Note that although some of the quantities are analytically solvable, the results from the simulation are used for purpose of consistency. In the tables, both bias and S.E. are calculated as the values obtained from the simulated data divided by the corresponding true parameters (relative bias and standard error). It is clear that under all three missing-data processes the QGPMLE is at least as good as the PMLE in terms of standard error and mean squared error, both

of which are essentially unbiased. The gain in efficiency of the QGPMLE over the PMLE is obvious. On the other hand, the MLE1 always has less bias and MSE when compared to the CC. As expected, under the MAR and MNAR, the CC is seriously biased. Note also that the MLE1 is seriously biased for β under the MNAR. The MLE1 has slightly larger standard errors than the MLE2 for α under MAR since it ignores the missing-data process that is dependent on α. The WEE is basically unbiased under all three missing-data mechanisms. However, its standard errors are larger than other approaches most of the time. Part of the reason is that the weighted estimating equation used is not the optimal one (Robins & Rotnitzky, 1995) since it would require knowledge of the exact distributional form, which is assumed to be unknown due to the semi-parametric nature of the WEE. However, as Robins & Rotnitzky (1995) showed in their simulation study, the improvement in efficiency of the optimal estimating equation relative to the general one is around 10-15%. Therefore, it appears that the WEE have larger standard errors than the QGPMLE. In general, the coverage probabilities for the WEE tend to be a little lower.

Regarding the estimation of α, the QGPMLE is at least as good as the MLE1 in terms of bias and standard error. The standard errors of the WEE are always larger than that of the QGPMLE and sometimes they are more than twice as large. Moreover, the QGPMLE essentially has the same efficiency as the MLE2. For the estimation of β, the WEE again has larger standard errors than the QGPMLE. The QGPMLE has a slightly larger standard error than the MLE2, which could come from lack of information of the distributional form. However, given the semi-parametric nature of the QGPMLE its performance is quite satisfactory. The QGPMLE also has reasonably small bias, which decreases as sample size increases. Moreover, the coverage probability of the QGPMLE is good for reasonable sample size.

In summary, the QGPMLE enjoys both distribution-free and easy-to-compute properties. Its performance is very close to that of the MLE that correctly specifies the missing-data process. Compared with the WEE, the QGPMLE is much more efficient. Since sometimes it is easier to model the mean structure for each pattern than to model the missing-data process, the QGPMLE provides a good alternative to the WEE. Even if it is hard to model the variance-covariance structure for each pattern under certain situations, the PMLE is

another alternative whose performance is similar to the WEE as can be seen from Tables 4.1 and 4.2.

4.2.2 POWER OF THE TEST OF MODEL MISSPECIFICATION

In Section 4.1.2, a Hausman (1978) type test to check model adequacy was proposed. The purpose of introducing this test is that it will be used in the construction of a parsimonious model. More specifically, it will be used to test whether or not some patterns share the same regression parameters though the shape of the distributional form is different from pattern to pattern. In this section, a small simulation study is presented to compare the power of this test under different distributional assumptions.

Consider a random bivariate vector (X, Y) such that the conditional distribution of Y given x, $[Y|x]$, is a mixture of two models with the first model having probability p of being selected to generate y. The two models are of the form:

$$\begin{aligned} \text{Model 1: } & E(Y|x) = \beta_1 x \quad Var(Y|x) = e^{|\beta_1 x|}, \\ \text{Model 2: } & E(Y|x) = \beta_2 x \quad Var(Y|x) = e^{|\beta_2 x|}. \end{aligned}$$

Moreover, suppose that there are three possible distributional forms for model 1 and model 2: (i) Normal Distribution (NOR), (ii) Double Exponential Distribution (DEP) and (iii) Triangle Distribution (TRI). The shapes of the three distributional forms are shown in Fig 4.1. Note that under all three distributional shapes, the mean and variance entirely determine the density function. It can be seen that under both of the models, the conditional variance of Y given x has an exponential relationship with the absolute value of the conditional mean. Therefore there is a substantial amount of variation that makes it relatively difficult to compare β_1 and β_2. Moreover, it is assumed that one does not know from which model each data point is drawn, which makes the problem even more complicated. Later we will see that the test we proposed in Section 4.1.2 can be used to detect the difference between β_1 and β_2.

Consider the null hypothesis $H_0 : \beta_1 = \beta_2$ versus $H_A : \beta_1 \neq \beta_2$. Throughout the simulation X follows a standard normal distribution and $p = 0.5$. Three combinations of

values of β_1 and β_2 are considered: (i) $\beta_1 = \beta_2 = -1$, (ii) $\beta_1 = -1$, $\beta_2 = 1$ and (iii) $\beta_1 = -2$, $\beta_2 = 2$. Six combinations of the distributional forms are considered: (a) (NOR, NOR), (b) (NOR, DEP), (c) (NOR, TRI), (d) (DEP, DEP), (e) (DEP, TRI) and (f) (TRI, TRI). Here the first entry refers to the shape for the first model and the second entry refers to the shape for the second model. Thus (NOR, DEP) indicates that the first model is a normal model and the second model is a double exponential model and so on.

The simulation results based on 1000 replicates are shown in Table 4.3. It can be seen that the power is quite similar among the six combinations of distributional shapes, particularly when the sample size reaches 500. Therefore the exact distributional form seems not to have a large influence on the power of this test. Moreover, the gain in power and the size of the test are limited as sample size increases from 200 to 500, which suggests that the converge rate might be relatively slow after the sample size increases beyond 200. It should be emphasized that lack of information on which model each of the data points is drawn from greatly limits the statistical power to detect the difference between β_1 and β_2. In this regard, lack of information on the exact distributional form has less influence. Another observation is that the difference between β_1 and β_2 seems to have a greater influence on power than the sample size does within the simulation setting.

Generally speaking, the power of the proposed test in Theorem 3 is hard to conjecture when applied to complicated models since many factors are involved. Since the test is most powerful for large difference in parameters as seen in Table 4.3, the idea is to sacrifice a little bias to achieve improvement in efficiency.

4.3 APPLICATION TO THE MINI MENTAL STATE EXAM (MMSE)

The proposed method was applied to a data set from the MoVIES. In the analysis data from waves 1 to 5 are included with intermittent missing cases and subjects whose outcomes were not observed at baseline excluded. The dataset is then composed of 1323 subjects, among which 271(20.5%), 164(12.4%), 144(10.9%), 155(11.7%) leave the study at wave 2 ($R = 1$), 3 ($R = 2$), 4 ($R = 3$) and 5 ($R = 4$), respectively, and 589(44.5%) are completers ($R = 5$) (see Table 4.4).

The response variable in our example is a neuropsychological test score called the Mini Mental State Exam (Folstein et al., 1975), or MMSE, which measures global cognitive performance and has generally been used as part of a screening battery to detect dementia. The MMSE was administered at 5 separate time points and we are interested in the score trajectory. The empirical trajectories of the mean MMSE over the five waves for different patterns are shown in Fig 4.2. Fig 4.2 clearly demonstrates that the score at baseline and the slope are different across patterns. Particularly, subjects who stay longer in the study tend to have a higher baseline score. It appears that the slopes for $R = 2, 3$ and 4 are quite similar, and are steeper than that of $R = 5$. From the observed data, it seems a linear model is a good approximation to describe the trajectory. Since no information is available with respect to the score beyond the dropout time, extrapolation based on the linear model is used to describe the trajectory after the dropout time. This certainly is not the strategy to fully explore a pattern-mixture model, which usually requires sensitivity analysis assuming different model structures regarding the under-identifiability issue. Here the purpose is to demonstrate the idea of simplification of identifiable pattern-mixture models and try to avoid other complications that might distract from the key point we are trying to make.

The dimension of the parameter space associated with the mean configuration in a pattern-mixture model is quite flexible, depending on the assumptions made. Several different linear models are chosen for the analysis of this data, assuming dropout is related to the unobserved MMSE score. Additionally, since the PMLE and QGPMLE lend themselves to comparison across models using the proposed test statistic given in Theorem 3, we also consider three variance structures for modelling this data. Let σ_{ij}^r denote the covariance between the ith and jth outcomes for a subject in pattern r. The three variance structures (all of them are independent of covariates) that will be considered are:

(i) $\sigma_{ij}^{r_1} = \sigma_{ij}^{r_2}$ for any r_1 ,$r_2 \geq max(i, j)$. For example σ_{23}^r is the same for all $r = 3, 4, 5$.

(ii) Let $A_1 = \{1, 2, 3\}$ and $A_2 = \{4, 5\}$, then (i) holds for any r_1, r_2 both in A_1 or both in A_2 and does not hold otherwise.

(iii) $\sigma_{ij}^{r_1} \neq \sigma_{ij}^{r_2}$ for any r_1 ,$r_2 \geq max(i, j)$ unless $r_1 = r_2$.

In other words, (i) requires that all estimable covariance parameters are the same across the missing-data patterns, (iii) specifies the structure in exactly the opposite way and (ii)

requires that the estimable covariance parameters be the same within the two groups of missing-data patterns and different between the two groups.

Both the PMLE and QGPMLE are calculated from pseudo likelihood functions based on the (multivariate) normal distribution. To obtain the PMLE, an identity covariance matrix is assumed for each pattern. For the QGPMLE, a consistent estimator of the covariance matrix is obtained for each pattern by averaging the residuals (calculated based on the PMLE) under each of the three covariance structures described above. Following covariates are included: age at baseline (*age*), sex (*female*: 1-female, 0-male), education level (*highedu*: 1-high school or higher, 0-otherwise) and time (t) from baseline in years.

To obtain information regarding how to simplify models, a saturated pattern-mixture model was first fitted to the data in which parameters from each pattern are set to be distinct except that the slope of t for pattern $R = 1$ is set to be the same as that of pattern $R = 2$, assuming variance structure (iii). The estimates and standard errors are shown in Table 4.5. From Table 4.5 it can be seen that some estimates across patterns are quite different (*female* for pattern $R = 3$ and $R = 4$) and some are very close (*intercept* for pattern $R = 2$ and $R = 3$). Then we try different pattern-mixture models (slope of t for $R = 1$ is always set to be equal to that of $R = 2$) to allow certain patterns to share certain parameters according to the observations from Table 4.5. Moreover, the test in Theorem 3 is used to check the adequacy of the different models in explaining the data. Several simplified models are obtained that seem to explain the data well. Furthermore, all of them give similar marginal parameter estimates. We pick one model which we will call the parsimonious pattern-mixture model (PPM) and compare it with the complete case analysis (CC), the observed data analysis (OD) and the saturated pattern-mixture model (SPM). Both the CC and OD are based on pseudo maximum likelihood estimation, ignoring the missing-data process. The PPM ($\chi^2(9) = 10.5$, $p = 0.31$) assumes variance structure (i) and specifies the mean structure to be

$$\begin{aligned} E(\text{MMSE}_{ij}) \;=\; & \beta_{01} * I(R_i = 1) + \beta_{023} * I(R_i = 2 \text{ or } R_i = 3) + \beta_{045} * I(R_i = 4 \text{ or } R_i = 5) \\ & + \beta_{t5} * I(R_i = 5) * t_{ij} + \beta_{t1234} * I(R_i < 5) * t_{ij} + \beta_{f3} * I(R_i = 3) * female_i \\ & + \beta_{f1245} * I(R_i \neq 3) * female_i + \beta_h * highedu_i + \beta_a * age_i, \end{aligned}$$

where i and j are subject index and wave index, respectively. Essentially the PPM says: (i) the effects of age at baseline and education level are the same across missing-data patterns; (ii) intercepts of the missing-data patterns form three distinct groups: $(R = 1)$, $(R = 2,$ $R = 3)$ and $(R = 4,\ R = 5)$; (iii) slopes of time are different for completers $(R = 5)$ and other patterns and (iv) sex effect is different for pattern $R = 3$ and other patterns. This model reflects some observations made from Table 3 and Fig 4.2. For example, the effect of gender for $R = 3$ is quite different from that of other patterns and completers have a less steep decline (Fig 4.2) than other patterns. We also fit larger models to take into account more subtle differences in parameters across patterns (e.g. sex effect can form 3 distinct groups: $(R = 1,\ R = 2,\ R = 4)$, $(R = 3)$ and $(R = 5)$). It turns out that these models give marginal parameter estimates that are similar to those obtained from the PPM. Moreover, although the effect of age at baseline seems to be heterogenous across patterns as can be seen from the SPM ($p = 0.09$ for Wald test of equality of age effects across patterns), models that distinguish it yield similar pooled parameter estimates as the PPM. Empirical calculation shows that some elements in the variance-covariance matrix across patterns are quite close whereas some are not. However, as long as the mean structure of the PPM is correctly specified, the estimates are still consistent even the variance-covariance structure deviates from the true one. The parameter estimates from the PPM are shown in Table 4.6.

In Table 4.7 we compare the marginal parameter estimates of the PPM and SPM (empirical proportion of each pattern as the weight) with parameter estimates of the CC and OD. From all four estimation procedures we see that females and subjects with a high school education or higher tend to have higher test scores and older subjects at baseline tend to have lower test scores. Clearly, the CC overestimates the baseline score since it is based on completers, who are much healthier than the rest of the population. It is also seen that both the CC and OD underestimate the decline rate since they do not take into account those test scores that would have been observed had the subjects not left the study. As a matter of fact, the test of model adequacy rejects the OD ($\chi^2(5) > 100$ for all three variance structures, $p < 0.001$), which indicates that MCAR assumption almost surely does not hold. The OD and the SPM can be thought as two extreme points of pattern-mixture models with the OD being the "smallest" pattern-mixture model and the SPM being the "largest" one.

The PPM is an "intermediate" model (most estimates fall between that of the OD and the SPM except the variable *female*). Although the efficiency improvement of PPM relative to the SPM is not substantial for most pooled parameters, the gain in efficiency for the slope of time is quite apparent. Thus the proposed approach allows one to balance bias and efficiency by fitting models with different complexity.

4.4 DISCUSSION

In this chapter a pseudo maximum likelihood approach is proposed for the estimation of parameters in a pattern-mixture model. Although analyses based on generalized linear models have been a major tool for non-Gaussian longitudinal data, it is often hard to justify the distributional assumptions for each missing-data pattern due to the complexity and limited information of the missing-data process. The theory of pseudo maximum likelihood estimation guarantees consistent estimators by assuming an exponential family even though the true distribution might not belong to this class, provided the mean structure is correctly specified. The work in this chapter is a direct extension of pseudo maximum likelihood estimation by applying the theory to the problem of estimating parameters based on more than one data generation mechanisms, for which monotone missing data present a typical example. Other semi-parametric methods include Robins et al. (1994, 1995) and Rotnitzky et al. (1998), who developed an approach for nonresponses in the framework of weighted estimating equations.

Another point we are trying to convey in this chapter involves efficiency consideration. We want to emphasize that although obtaining an unbiased estimator has been the main goal for most research on missing data, the gain in practical applications is often limited because the missing-data process is poorly understood most of the time. On the other hand, there is still room for improvement in efficiency, which is also important since we already lose efficiency due to nonresponse. Sensitivity analysis has been an important strategy to explore different assumptions regarding the unverifiable elements in a pattern-mixture model for a non-ignorable missing outcome. The approach described in this chapter can be thought of as a supplement to such analysis in the hope to balance unbiasness and efficiency.

For the example in Section 4.3, the covariance components are independent of the covariates. An alternative is to assume a linear mixed effects Gaussian model for our data (e.g. with *intercept* and t as the random effects). Gourieroux et al. (1984) introduced a consistent estimator based on the quadratic exponential family, among which is the multivariate normal distribution. Then a simultaneous estimator of all the parameters involved (both parameters of mean structure and variance structure) by maximizing the pseudo likelihood function based on the multivariate normal distribution is consistent and asymptotically normal, provided the mean and variance structure are correctly specified. However, this will need a numerical algorithm and thus lose the advantage of an explicit solution. Since the time interval between two consecutive waves in our example is approximately two years for each subject, it is reasonable to assume a homogeneous covariance structure.

Table 4.1: Simulation results (50% missing) for complete case analysis (CC), MLE ignoring the missing-data mechanism (MLE1), the weighted estimating equation (WEE) method, the PMLE, the QGPMLE and the MLE based on correct specification of the missing-data process (MLE2). Three types of missingness simulated are covariate dependent missing (CDM), missing at random (MAR) and missing not at random (MNAR)

Sample size	Bias×100 S.E.×100 $\sqrt{MSE}$×100 95% cover prb	CC			MLE1			WEE			PMLE			QGPMLE			MLE2		
		CDM	MAR	MNAR	CDM	MAR	MNAR	CDM	MAR	MNAR	CDM	MAR	MNAR	CDM	MAR	MNAR	CDM	MAR	MNAR
N=100	α	0	-50	-29	0	0	0	0	-1	0	0	0	0	0	0	0	0	0	0
		14	7	10	10	10	10	14	19	16	13	12	14	10	8	10	10	8	10
		14	50	31	10	10	10	14	19	16	13	12	14	10	8	10	10	8	10
		0.947	0.000	0.232	0.957	0.957	0.957	0.933	0.898	0.932	0.941	0.929	0.924	0.957	0.944	0.953	0.957	0.943	0.945
	β	1	-50	-50	0	0	-29	1	-1	0	1	0	2	0	-2	-2	0	0	0
		20	10	10	17	17	12	28	42	43	28	33	34	22	21	22	17	17	16
		20	51	51	17	17	31	28	42	43	28	33	34	22	21	22	17	17	16
		0.947	0.032	0.025	0.942	0.939	0.349	0.919	0.889	0.899	0.869	0.851	0.860	0.928	0.921	0.921	0.942	0.928	0.944
N=200	α	0	-50	-29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		10	5	8	8	8	8	10	14	12	10	10	10	8	6	7	8	6	7
		10	50	30	8	8	8	10	14	12	10	10	10	8	6	7	8	6	7
		0.946	0.000	0.047	0.944	0.944	0.944	0.942	0.887	0.907	0.944	0.921	0.936	0.944	0.941	0.946	0.944	0.939	0.939
	β	1	-50	-50	0	0	-29	0	0	0	0	0	1	0	0	1	0	0	0
		14	7	7	12	12	8	21	31	31	20	23	22	16	15	15	12	12	12
		14	50	50	12	12	30	21	31	31	20	23	22	16	15	15	12	12	12
		0.934	0.000	0.000	0.942	0.942	0.119	0.936	0.900	0.909	0.942	0.899	0.904	0.942	0.935	0.944	0.942	0.948	0.940

Table 4.2: Simulation results (25% missing) for complete case analysis (CC), MLE ignoring the missing-data mechanism (MLE1), the weighted estimating equation (WEE) method, the PMLE, the QGPMLE and the MLE based on correct specification of the missing-data process (MLE2). Three types of missingness simulated are covariate dependent missing (CDM), missing at random (MAR) and missing not at random (MNAR)

Sample size	Bias×100 S.E.×100 $\sqrt{MSE}$×100 95% cover prb	CC			MLE1			WEE			PMLE			QGPMLE			MLE2		
		CDM	MAR	MNAR	CDM	MAR	MNAR	CDM	MAR	MNAR	CDM	MAR	MNAR	CDM	MAR	MNAR	CDM	MAR	MNAR
N=100	α	0	-26	-14	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
		12	9	10	10	10	10	13	16	12	14	14	14	10	9	10	10	8	10
		12	28	17	10	10	10	13	16	12	14	14	14	10	9	10	10	8	10
		0.941	0.229	0.668	0.947	0.947	0.947	0.956	0.895	0.927	0.916	0.918	0.917	0.947	0.949	0.947	0.947	0.959	0.949
	β	0	-26	-25	0	0	-14	0	0	0	0	-1	-1	0	-1	-1	0	0	0
		17	12	12	16	15	13	22	27	27	26	28	28	20	18	20	16	15	16
		17	29	28	16	15	19	22	27	27	26	28	28	20	18	20	16	15	16
		0.925	0.439	0.448	0.932	0.930	0.744	0.932	0.942	0.920	0.867	0.858	0.852	0.934	0.927	0.913	0.932	0.940	0.921
N=200	α	0	-25	-14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		8	6	7	7	7	7	9	12	8	10	10	10	7	6	7	7	6	7
		8	26	16	7	7	7	9	12	8	10	10	10	7	6	7	7	6	7
		0.936	0.054	0.511	0.950	0.950	0.950	0.973	0.914	0.950	0.931	0.908	0.929	0.950	0.938	0.944	0.950	0.949	0.940
	β	0	-25	-25	0	0	-13	0	0	-1	1	2	0	0	0	0	0	0	0
		12	9	9	11	11	10	16	21	18	18	20	20	14	14	14	11	10	11
		12	26	26	11	11	16	16	21	18	18	20	20	14	14	14	11	10	11
		0.938	0.233	0.233	0.944	0.941	0.669	0.960	0.930	0.946	0.889	0.901	0.889	0.925	0.934	0.936	0.944	0.943	0.942

Table 4.3: Power of testing $H_0 : \beta_1 = \beta_2$ vs. $H_A : \beta_1 \neq \beta_2$ for six combinations of distributional shapes (1000 replicates, $p = 0.5$)

Parameters	N	(NOR, NOR)	(NOR, DEP)	(NOR, TRI)	(DEP, DEP)	(DEP, TRI)	(TRI, TRI)
$\beta_1 = \beta_2 = -1$	100	0.072	0.061	0.068	0.078	0.084	0.080
	200	0.057	0.054	0.053	0.073	0.065	0.067
	500	0.046	0.058	0.057	0.061	0.058	0.062
$\beta_1 = -1,\ \beta_2 = 1$	100	0.481	0.464	0.474	0.485	0.498	0.510
	200	0.533	0.511	0.524	0.541	0.538	0.550
	500	0.556	0.558	0.558	0.545	0.562	0.563
$\beta_1 = -2,\ \beta_2 = 2$	100	0.725	0.696	0.705	0.715	0.725	0.738
	200	0.814	0.821	0.833	0.808	0.804	0.813
	500	0.853	0.853	0.850	0.860	0.855	0.850

Table 4.4: Distribution of the missing-data patterns of the MMSE from MoVIES data

R	Wave1	Wave2	Wave3	Wave4	Wave5	Frequency(%)
1	•	×	×	×	×	271 (20.5)
2	•	•	×	×	×	164 (12.4)
3	•	•	•	×	×	144 (10.9)
4	•	•	•	•	×	155 (11.7)
5	•	•	•	•	•	589 (44.5)
Total						1323 (100)
•: Observed, ×: Missing						

Table 4.5: Parameter estimates from the saturated pattern-mixture model (SPM) for the MMSE

	R=1		R=2		R=3		R=4		R=5	
Parameter	est	S.E.	est	S.E.	est	S.E.	est	S.E.	est	S.E.
intercept	25.328	0.454	25.917	0.346	25.945	0.324	26.977	0.267	26.717	0.116
t	-0.620	0.124	-0.620	0.124	-0.527	0.068	-0.387	0.048	-0.137	0.012
female	0.074	0.535	0.093	0.425	1.104	0.343	0.020	0.288	0.656	0.116
highedu	1.338	0.556	1.385	0.416	1.145	0.353	1.156	0.287	0.919	0.119
age	-0.207	0.039	-0.103	0.033	-0.148	0.030	-0.090	0.025	-0.096	0.013

Table 4.6: Parameter estimates from the parsimonious pattern-mixture model (PPM) for the MMSE

	R=1		R=2		R=3		R=4		R=5	
Parameter	est	S.E.	est	S.E.	est	S.E.	est	S.E.	est	S.E.
intercept	25.065	0.181	25.763	0.163	25.763	0.163	26.671	0.130	26.671	0.130
t	-0.488	0.030	-0.488	0.030	-0.488	0.030	-0.488	0.030	-0.155	0.016
female	0.349	0.121	0.349	0.121	1.337	0.262	0.349	0.121	0.349	0.121
highedu	1.233	0.118	1.233	0.118	1.233	0.118	1.233	0.118	1.233	0.118
age	-0.150	0.010	-0.150	0.010	-0.150	0.010	-0.150	0.010	-0.150	0.010

Table 4.7: Parameter (pooled parameter) estimates for the complete case analysis (CC), the observed data analysis (OD), the parsimonious pattern-mixture model (PPM) and the saturated pattern-mixture model (SPM)

	CC		OD		PPM		SPM	
Parameter	est (S.E.)	p value	est (S.E.)	p value	est (S.E.)	p value	est (S.E.)	p value
intercept	26.717 (0.116)	< 0.001	25.936 (0.113)	< 0.001	26.277 (0.112)	< 0.001	26.406 (0.107)	< 0.001
t	-0.137 (0.012)	< 0.001	-0.180 (0.015)	< 0.001	-0.302 (0.017)	< 0.001	-0.318 (0.025)	< 0.001
female	0.656 (0.116)	< 0.001	0.633 (0.118)	< 0.001	0.433 (0.116)	< 0.001	0.482 (0.120)	< 0.001
highedu	0.919 (0.119)	< 0.001	1.317 (0.121)	< 0.001	1.234 (0.118)	< 0.001	1.074 (0.122)	< 0.001
age	-0.096 (0.013)	< 0.001	-0.188 (0.010)	< 0.001	-0.150 (0.010)	< 0.001	-0.117 (0.011)	< 0.001

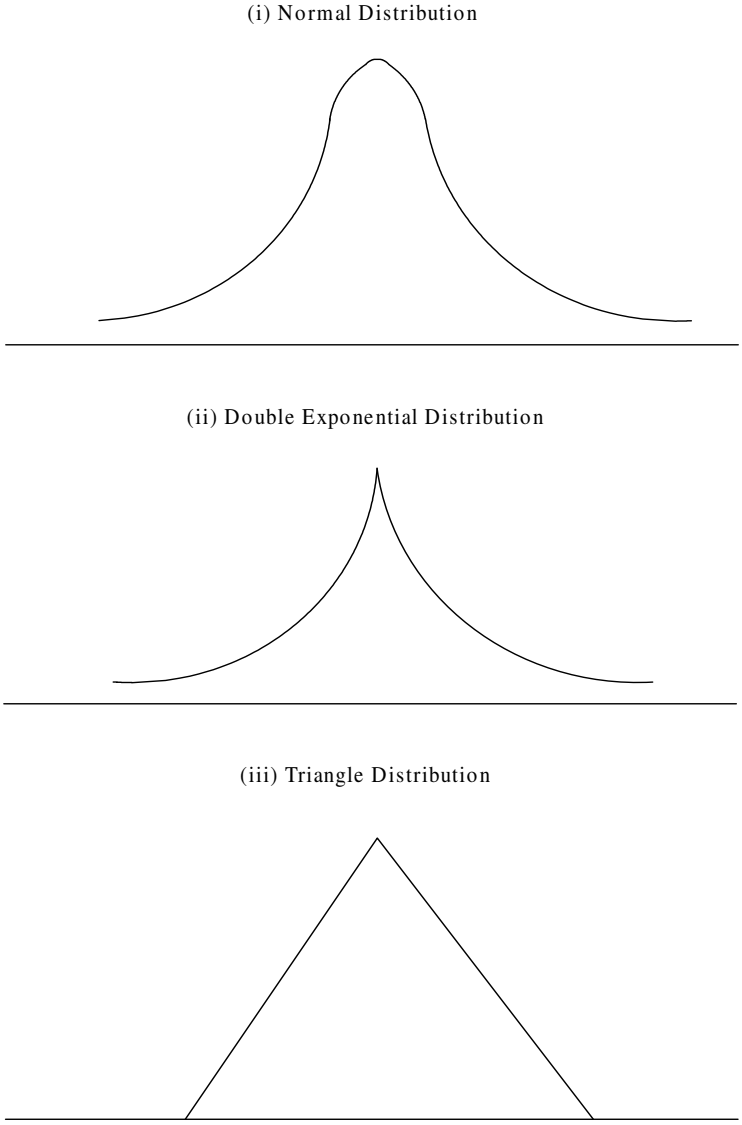

Figure 4.1: Three distribution forms

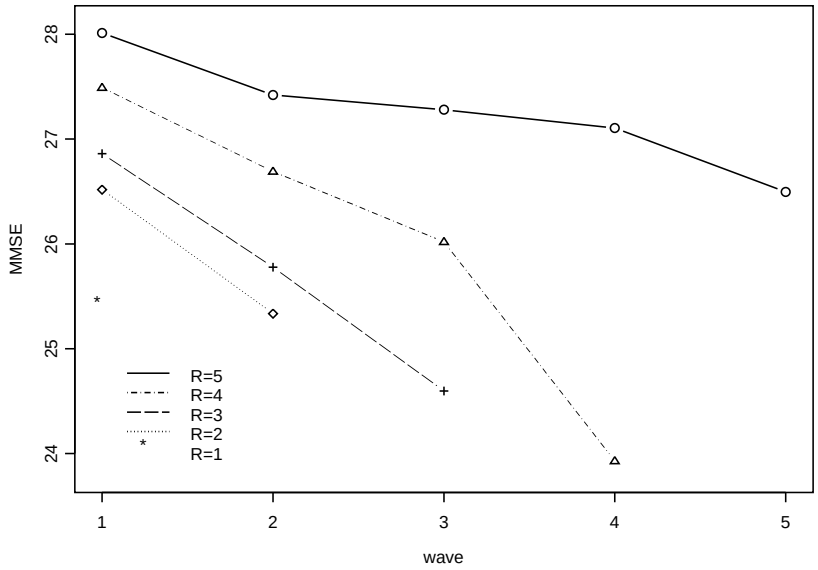

Figure 4.2: The mean MMSE scores over waves for $R = 1, 2, 3, 4$ and 5

5.0 CONCLUSIONS

A parametric selection model based on a multivariate normal copula and a semi-parametric pattern-mixture model are proposed for regression analysis with non-ignorable missing outcomes in longitudinal studies. The first approach connects the missing-data process and the outcome directly through a normal copula so that the correlation between them is reflected in the copula function. The second approach relaxes assumptions regarding the distributional form for each pattern in a pattern-mixture model and only requires specification of the first two moments. Since the two methods fall into selection models and pattern-mixture models, they share the same advantages and disadvantages of other models within the same class. The major contribution of the normal copula-based selection model is that it provides a framework to combine the outcomes and missing-data indicators together. With different copula functions, this method can be quite general and capable of handling different missing-data processes. Pattern-mixture models with pseudo maximum likelihood estimation allow one to include a more general class of distributions for each pattern. Moreover, the model misspecification test based on the PMLE and QGPMLE allows one to balance unbiasness and efficiency in model fitting.

A more general form of the copula-based selection models can be constructed through a linkage function (Li et al., 1996). A linkage function is a tool to construct multivariate distributions with given multivariate margins. Thus one can model $[Y|X]$ and $[R|X]$ separately and join them together by a linkage function. In that way, the dependence structure of Y and R given X is independent of $[Y|X]$ and $[R|X]$. This is a desirable property for modelling the joint distribution of the outcomes and the missing-data indicators since often we need $[Y|X]$ and $[R|X]$ to possess a specific dependence structure that is independent

of their multivariate margins. This method can be an extension of the model discussed in Chapter 3 for future investigation. For pattern-mixture models, Roy (in press) proposed a latent class-based pattern-mixture model, in which each pattern has different probabilities to fall into each latent class. Then the distribution of the response for a given pattern is a weighted sum of the distribution within each latent class. A linear random effects model is then built for each latent class. Therefore, as long as the data include complete cases, all parameters are identifiable. The semi-parametric method described in Chapter 4 can be applied to this setting directly. Moreover, simplification of a model in Chapter 4 is done by allowing different patterns to share the same regression parameters. A more general form is to allow parameters in a set of patterns to be a function (e.g. linear) of their counterparts in another set of patterns.

Due to the under-identification nature of models for non-ignorable missing outcomes, it is essentially impossible to verify assumptions made to build a model unless extra relevant information is available. Therefore, sensitivity analysis is of great significance in that it allows one to evaluate the variation of the results based on different assumptions. Our approaches provide a starting point to do such sensitivity analysis. For example, one can compare the results obtained from selection models built on different copula families or one can experiment with different simplified pattern-mixture models based on different assumptions as to the non-identifiable parameters. It is in this way that we might have a better understanding of the missing-data process involved.

APPENDIX A

PROOF OF THEOREM 1 AND 2

Consistency

I assume that the parameter space, Θ, is a Cartesian product of the subspaces of θ's elements. Assume also that Θ is compact. For all r, the $\theta(r)$'s are identifiable in the sense that $f_r(x, \theta_1(r)) = f_r(x, \theta_2(r))$ *a.s.* with respect to the distribution of $X|R = r$ implies $\theta_1(r) = \theta_2(r)$. Other regularity conditions referred to in the proof can be found in Gourieroux et al. (1984).

Let $\bar{l}_{n_r}(\theta(r)) = \sum_{j=1}^{n_r} \log[e_r(y_{rj}, f_r(x_{rj}, \theta(r))]/n_r$ and $\bar{l}_n(\theta) = \sum_{r=1}^{m} w_r \bar{l}_{n_r}$, where $w_r = n_r/n$. Moreover, $\phi_r(\theta(r)) = E\{\log[e_r(Y_{(r)}, f_r(X_{(r)}, \theta(r))]|R = r\}$. Thus under regularity conditions and due to the properties of the exponential family, we have the following facts for all r: ϕ_r has a unique maximum at $\theta_0(r)$ over Θ_r, where Θ_r is the parameter space of $\theta(r)$; $\bar{l}_{n_r}(\theta(r)) \xrightarrow{a.s.} \phi_r(\theta(r))$; $\hat{\theta}_{n_r} = \text{argmax}\bar{l}_{n_r}$ converges a.s. to $\theta_0(r)$ (Gourieroux et al., 1984).

Moreover, $\phi_r(\theta(r))$ is a continuous function of $\theta(r)$ and $\bar{l}_{n_r}(\theta(r))$ converges to $\phi_r(\theta(r))$ uniformly. These two conditions imply that $\forall\ \epsilon > 0$, for large enough n,

$$\bar{l}_n(\theta_0) = \sum w_r \bar{l}_{n_r}(\theta_0(r)) > \sum w_r \bar{l}_{n_r}(\hat{\theta}_{n_r}) - \epsilon.$$

Since $\hat{\theta}_n$ maximizes $\bar{l}_n$ and $\bar{l}_n(\theta_0)$ can be arbitrarily close to its upper bound, $\bar{l}_{n_r}(\hat{\theta}_n(r))$ has to become arbitrarily close to $\bar{l}_{n_r}(\hat{\theta}_{n_r})$ for large enough n and for all r. Again, due to the uniform convergence of $\bar{l}_{n_r}(.)$ and uniform continuity of $\phi_r(.)$, for large enough n, $|\phi_r(\hat{\theta}_n(r)) - \phi_r(\theta_0(r))| < \epsilon$ for all r.

Now suppose $\hat{\theta}_n = \text{argmax}\bar{l}_n(\theta)$ does not converge to θ_0. Then $\exists\ \delta > 0$ such that $|\hat{\theta}_n - \theta_0| \geq \delta$ infinitely often. Thus $|\hat{\theta}_n(r) - \theta_0(r)| \geq \delta/\sqrt{m}$ infinitely often for at least one r. Since Θ_r is compact, $\phi_r(.)$ assumes its maximum M_r on $\Theta_r^* = \Theta_r - N(\delta)$, where $N(\delta) = \{\theta(r) : |\theta(r) - \theta_0(r)| < \delta/\sqrt{m}\}$. Since ϕ_r has a unique maximum at $\theta_0(r)$ over Θ_r, $M_r < \phi_r(\theta_0(r))$. However, there is an infinite sequence $\hat{\theta}_n(r)$ on Θ_r^* such that $\phi_r(\theta_0(r))$ is the supremum of the corresponding sequence $\phi_r(\hat{\theta}_n(r))$. Therefore $M_r = \phi_r(\theta_0(r))$. Contradiction. Thus $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$.

Asymptotic normality

Expand $\bar{l}'_n(\hat{\theta}_n)$ at θ_0: $0 = \bar{l}'_n(\hat{\theta}_n) = \bar{l}'_n(\theta_0) + \bar{l}''_n(\theta_0)(\hat{\theta}_n - \theta_0) + o_p(\hat{\theta}_n - \theta_0)$, resulting in

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = [-\bar{l}''_n(\theta_0)]^{-1}[\sqrt{n}\bar{l}'_n(\theta_0)] + o_p(1).$$

By the strong law of large numbers,

$$-\bar{l}''_n(\theta_0) \xrightarrow{a.s.} \sum_{r=1}^{m} b_r E[\frac{\partial f_r}{\partial \theta_0}(\Sigma_r^{\theta_0})^{-1}\frac{\partial f_r}{\partial \theta_0'}|R = r] = E[\frac{\partial f_R}{\partial \theta_0}(\Sigma_R^{\theta_0})^{-1}\frac{\partial f_R}{\partial \theta_0'}] = J,$$

where $b_r = \Pr[R = r]$.

Moreover, $\sqrt{n}\bar{l}'_n(\theta_0) = \sum_{r=1}^{m}\sqrt{n_r/n}\sqrt{n_r}\bar{l}'_{n_r}(\theta_0(r)) \xrightarrow{d} \sum_{r=1}^{m}\sqrt{b_r}Z_r$, where the Z_r's are independently distributed as $N(0, E[\frac{\partial f_r}{\partial \theta_0}(\Sigma_r^{\theta_0})^{-1}\Omega_r(\Sigma_r^{\theta_0})^{-1}\frac{\partial f_r}{\partial \theta_0'}]|R = r])$. Therefore we have

$$\sqrt{n}\bar{l}'_n(\theta_0) \xrightarrow{d} N(0, I).$$

Then the asymptotic covariance matrix of the PMLE in Theorem 1 follows.

According to Burguete et al.'s (1982) results, the asymptotic covariance matrix of the QGPMLE can be calculated as if the nuisance parameter was known. Then the asymptotic covariance matrix of the QGPMLE follows directly. The proof that this covariance matrix is the lower bound of that in Theorem 1 is exactly the same as what was shown in Gourieroux et al. (1984).

APPENDIX B

PROOF OF THEOREM 3

Here I show the proof when there is only one pattern. The same reasoning also applies to general situations. In this setting, i is the subject index.

Let

$$\begin{aligned}\lambda_n(\theta) &= \sum_{i=1}^{n} \log[e(y_i, f(x_i, \theta))]/n, \\ \psi_n(\theta) &= \sum_{i=1}^{n} \log[e^*(y_i, f(x_i, \theta), g(f(x_i, \breve{\theta}_n), h(x_i, \breve{\alpha}_n)))]/n.\end{aligned}$$

Here $\breve{\theta}_n$ and $\breve{\alpha}_n$ are strongly $\sqrt{n}$-consistent estimators.

By Taylor expansion:

$$\begin{aligned}0 = \frac{\partial \lambda_n(\hat{\theta}_n)}{\partial \theta} &= \frac{\partial \lambda_n(\theta_0)}{\partial \theta} + \frac{\partial^2 \lambda_n(\theta_*)}{\partial \theta \partial \theta'}(\hat{\theta}_n - \theta_0), \\ 0 = \frac{\partial \psi_n(\tilde{\theta}_n)}{\partial \theta} &= \frac{\partial \psi_n(\theta_0)}{\partial \theta} + \frac{\partial^2 \psi_n(\theta_{**})}{\partial \theta \partial \theta'}(\tilde{\theta}_n - \theta_0).\end{aligned}$$

Here θ_* and θ_{**} are points on the segment that connects θ_0 with $\hat{\theta}_n$ and $\tilde{\theta}_n$.

Let $l = \log e$, $l^* = \log e^*$ and define

$$\begin{aligned}A(\theta_0) &= E\left\{-\frac{\partial^2 l(Y, f(X, \theta_0)}{\partial \theta \partial \theta'}\right\}, \\ B(\theta_0) &= E\left\{\frac{\partial l(Y, f(x, \theta_0)}{\partial \theta}\frac{\partial l(Y, f(x, \theta_0)}{\partial \theta'}\right\}, \\ C(\theta_0) &= E\left\{-\frac{\partial^2 l^*(Y, f(X, \theta_0)}{\partial \theta \partial \theta'}\right\}, \\ D(\theta_0) &= E\left\{\frac{\partial l^*(Y, f(x, \theta_0)}{\partial \theta}\frac{\partial l^*(Y, f(x, \theta_0)}{\partial \theta'}\right\}.\end{aligned}$$

Then it is easy to show that:

$$\sqrt{n}\begin{pmatrix}\hat{\theta}_n-\theta_0\\ \tilde{\theta}_n-\theta_0\end{pmatrix}\xrightarrow{d}N(0,V),\ V=\begin{pmatrix}V_{11}&V_{12}\\ V_{21}&V_{22}\end{pmatrix},$$

where

$$\begin{aligned}V_{11}=A(\theta_0)^{-1}B(\theta_0)A(\theta_0)^{-1},\ V_{12}&=A(\theta_0)^{-1}K(\theta_0)C(\theta_0)^{-1},\\ V_{21}=V_{12}',\ V_{22}&=C(\theta_0)^{-1}D(\theta_0)C(\theta_0)^{-1},\end{aligned}$$

and

$$K(\theta_0)=E\left\{\frac{\partial l(Y,f(x,\theta_0)}{\partial\theta}\frac{\partial l^*(Y,f(x,\theta_0)}{\partial\theta'}\right\}.$$

After some algebra, it can be shown that $D(\theta_0)=C(\theta_0)$ and $K(\theta_0)=A(\theta_0)$. Then $\sqrt{n}(\hat{\theta}_n-\tilde{\theta}_n)\xrightarrow{d}N(0,G)$, where $G=V_{11}+V_{22}-V_{12}-V_{21}=V_{11}-V_{22}=V_h(\theta_0)-V_t(\theta_0)$. Therefore the result of Theorem 3 follows.

BIBLIOGRAPHY

Amemiya, T. (1984). Tobit models: a survey. *Journal of Econometrics* **24**, 3–61.

Burguete, J. F., Gallant, A. R. & Souza, G. (1982). On unification of the asymptotic theory of nonlinear econometric models. *Econometric Reviews* **1**, 151–190.

Conaway, M. R. (1992). The analysis of repeated categorical measurements subject to nonignorable nonresponse. *Journal of the American Statistical Association* **87**, 817–824.

Conaway, M. R. (1993). Non-ignorable non-response models for time-ordered categorical variables. *Appl. Statist.* **42**, 105–115.

Diggle, P. & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Appl. Statist.* **43**, 49–93.

Fitzmaurice, G. M. & Laird, N. M. (2000). Generalized linear mixture models for handling nonignorable dropouts in longitudinal studies. *Biostatistics* **1**, 141–156.

Follmann, D. & Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics* **51**, 151–168.

Folstein, M. F., Folstein, S. E. & McHugh, P. R. (1975). Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research* **12**, 189–198.

Gay, D. M. (1983). Algorithm 611: subroutines for unconstrained minimization using a model/trust-region approach. *ACM transactions on mathematical software* **9**, 503–524.

Gong, G. & Samaniego, F. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics* **9**, 861–869.

Gourieroux, C., Monfort, A. & Trognon, A. (1984). Pseudo maximum likelihood methods: theory. *Econometrica* **52**, 681–700.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica* **46**, 1251–1272.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* **5**, 475–492.

Hogan, J. W. & Laird, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine* **16**, 239–258.

Ibrahim, J. G., Chen, M.-H. & Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika* **88**, 551–564.

Kenward, M. G., Molenberghs, G. & Thijs, H. (2003). Pattern-mixture models with proper time dependence. *Biometrika* **90**, 53–71.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79–86.

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine* **7**, 305–315.

Lezak, M. D. (1995). *Neuropsychological assessment*. New York: Oxford University Press, 3rd ed.

Li, H., Scarsini, M. & Shaked, M. (1996). Linkages: a tool for the construction of multivariate distributions with given nonoverlapping multivariate marginals. *Journal of Multivariate Analysis* **56**, 20–41.

Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lipsitz, S. R., Laird, N. M. & Harrington, D. P. (1992). A three-stage estimator for studies with repeated and possibly missing binary outcomes. *Appl. Statist.* **41**, 203–213.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* **88**, 125–134.

Little, R. J. A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika* **81**, 471–483.

Little, R. J. A. (1995). Modeling drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* **90**, 1112–1121.

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

Little, R. J. A. & Wang, Y. (1996). Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics* **52**, 98–111.

Meester, R. B. & Mackay, J. (1994). A parametric model for clustered correlated categorical data. *Biometrics* **50**, 954–963.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimisation. *Computer Journal* **7**, 303–313.

Nelsen, R. B. (1998). *An introduction to copulas*. New York: Springer.

Park, T. & Lee, S. (1999). Simple pattern-mixture models for longitudinal data with missing observations: analysis of urinary incontinence data. *Statistics in Medicine* **18**, 2933–2941.

Parke, W. (1986). Pseudo maximum likelihood estimation: The asymptotic distribution. *The Annals of Statistics* **14**, 355–357.

Robins, J. M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine* **16**, 21–37.

Robins, J. M. & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association* **90**, 122–129.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90**, 106–121.

Rotnitzky, A., Robins, J. M. & Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* **93**, 1321–1339.

Roy, J. (in process). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics* .

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Jouranl of the American Statistical Association* **91**, 473–489.

Sklar, A. (1959). Fonctions de repartition a n dimensions et leurs marges. *Publications De L'institut De Statistique De L'universite De Paris* **8**, 229–231.

Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G. & Curran, D. (2002). Strategies to fit pattern-mixture models. *Biostatistics* **3**, 245–265.

Troxel, A. B., Harrington, D. P. & Lipsitz, S. R. (1998a). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Appl. Statist.* **47**, 425–438.

Troxel, A. B., Lipsitz, S. R. & Harrington, D. P. (1998b). Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data. *Biometrika* **85**, 661–672.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika* **61**, 439–447.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association* **76**, 419–433.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Wu, M. C. & Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44**, 175–188.