

**LATENT VARIABLE MODELS FOR  
LONGITUDINAL STUDY WITH INFORMATIVE  
MISSINGNESS**

by

**Li Qin**

B.S. in Biology,

University of Science and Technology of China, 1998

M.S. in Statistics,

North Dakota State University, 2001

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**

University of Pittsburgh

2006

UNIVERSITY OF PITTSBURGH  
GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Li Qin

It was defended on

April 4th, 2006

and approved by

Lisa A. Weissfeld, Ph.D., Professor, Department of Biostatistics, Graduate School of Public  
Health, University of Pittsburgh

Sati Mazumdar, Ph.D., Professor, Department of Biostatistics, Graduate School of Public  
Health, University of Pittsburgh

Stewart Anderson, Ph.D., Associate Professor, Department of Biostatistics, Graduate  
School of Public Health, University of Pittsburgh

Michele D. Levine, Ph.D., Assistant Professor, Department of Psychiatry, School of  
Medicine, University of Pittsburgh

Dissertation Director: Lisa A. Weissfeld, Ph.D., Professor, Department of Biostatistics,  
Graduate School of Public Health, University of Pittsburgh

Copyright © by Li Qin  
2006

# LATENT VARIABLE MODELS FOR LONGITUDINAL STUDY WITH INFORMATIVE MISSINGNESS

Li Qin, PhD

University of Pittsburgh, 2006

Missing problem is very common in today's public health studies because of responses measured longitudinally. In this dissertation we proposed two latent variable models for longitudinal data with informative missingness. In the first approach, a latent variable model is developed for the categorical data, dividing the observed data into two latent classes: a 'regular' class and a 'special' class. Outcomes belonging to the regular class can be modeled using logistic regression and the outcomes in the special class have pre-deterministic values. Under the important assumption of conditional independence in the latent variable models, the longitudinal responses and the missingness process are independent given the latent classes. Parameters that we are interested in are estimated by the method of maximum likelihood based on the above assumption and correlation between responses. In the second approach, the latent variable in the proposed model is continuous and assumed to be normally distributed with unity variance. In the latent variable model, the values of the latent variable are affected by the missing patterns and the latent variable is also a covariate in modeling the longitudinal responses. We use the EM algorithm to obtain the estimates of the parameters and Gauss-Hermite quadrature is used to approximate the integral of the latent variable. The covariance matrix of the estimates can be calculated by using the bootstrap method or obtained from the inverse of the Fisher information matrix of the final marginal likelihood.

## PREFACE

This dissertation is organized in the following way. In Chapter 1, we give an introduction on missing mechanisms, methods for dealing with missing data, especially the latent variable models, and a simple description of the whole dissertation. In Chapter 2, we present literature reviews on approaches for longitudinal missing data and latent variable models. In Chapter 3, we propose a latent class model for the categorical outcomes and compare it with the weighted GEE and the shared parameter model in the simulation and the application. In Chapter 4, a latent variable model is developed for the complicated intermittent missing data in which the continuous latent variable link the missingness process and the longitudinal component. We give the summary of this dissertation in the final chapter.

I wish to thank my advisor, Dr. Lisa Weissfeld, for her guidance and constant support in this research. Her suggestions and careful corrections of this dissertation help me a lot. I would like to thank Dr. Stewart Anderson, Dr. Sati Mazumdar and Dr. Michele Levine for serving as my committee members and their valuable discussions and suggestions of my proposal and this dissertation. I would like to express my appreciation to Dr. Michele Levine for providing one of the data sets in this dissertation.

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

## TABLE OF CONTENTS

<b>PREFACE</b> . . . . .	v
<b>1.0 INTRODUCTION</b> . . . . .	1
<b>2.0 LITERATURE REVIEW FOR MISSING DATA ANALYSIS AND LATENT VARIABLE MODELS</b> . . . . .	7
2.1 APPROACHES FOR LONGITUDINAL MISSING DATA . . . . .	7
2.1.1 GENERALIZED ESTIMATING EQUATIONS (GEE) . . . . .	7
2.1.2 LIKELIHOOD-BASED METHODS . . . . .	8
2.1.2.1 SHARED PARAMETER MODELS . . . . .	9
2.1.2.2 TRANSITION MODELS . . . . .	10
2.1.2.3 OTHER METHODS . . . . .	11
2.2 LATENT VARIABLE MODELS . . . . .	12
<b>3.0 A LATENT CLASS MODEL FOR LONGITUDINAL BINARY RESPONSES WITH INFORMATIVE DROPOUT</b> . . . . .	16
3.1 INTRODUCTION . . . . .	16
3.2 MODELS . . . . .	19
3.2.1 Latent Class Model (LCM) . . . . .	19
3.2.2 Shared Parameter Model (SPM) . . . . .	22
3.2.3 Weighted GEE (WGEE) . . . . .	23
3.3 SIMULATION RESULTS . . . . .	23
3.4 APPLICATION TO THE SMOKING CESSATION STUDY . . . . .	26
3.5 DISCUSSION . . . . .	28

<b>4.0 AN EXTENSION OF LATENT VARIABLE MODEL FOR INFORMATIVE INTERMITTENT MISSING DATA . . . . .</b>	<b>32</b>
4.1 INTRODUCTION . . . . .	32
4.2 MODEL SPECIFICATION AND ESTIMATION . . . . .	35
4.2.1 MODEL SPECIFICATION . . . . .	35
4.2.2 ESTIMATION . . . . .	37
4.3 APPLICATION TO THE KIDQUEST DATA . . . . .	39
4.3.1 DATA DESCRIPTIONS AND MODEL SPECIFICATIONS . . . . .	39
4.3.2 ASSESSING FIT OF THE MODEL . . . . .	42
4.4 DISCUSSION . . . . .	42
<b>5.0 SUMMARY . . . . .</b>	<b>49</b>
<b>BIBLIOGRAPHY . . . . .</b>	<b>51</b>

## LIST OF TABLES

1	Four kinds of latent variable models. . . . .	4
2	Simulation results for weighted GEE (WGEE), shared parameter model (SPM) and latent class model (LCM) (sample size = 200, informative dropout). . . . .	29
3	Summary of outcomes for the smoking cessation study. . . . .	30
4	Marginal parameter estimates, estimated standard errors and Z-values for the smoking cessation study (Modelling Pr(abstinent)). . . . .	30
5	Estimates, estimated standard errors and Z-values for the latent classes, $e_1$ and $e_0$ under the proposed method for the smoking cessation study. . . . .	30
6	Hermite integration $\int_{-\infty}^{\infty} g(x)dx = \sum_{i=1}^n w_i e^{x_i^2} g(x_i)$ for $n = 10$ . . . . .	44
7	Descriptive statistics for $Z_{BMI}$ . . . . .	44
8	Descriptive statistics for missingness (frequency and percentage in the table are for the missingness). . . . .	44
9	Distribution of the missing patterns for KIDQUEST data . . . . .	45
10	Estimates, estimated standard errors and Z-values for parameter of latent distribution . . . . .	45
11	Estimates, estimated standard errors and Z-values for modelling the outcomes, $Z_{BMI}$ . . . . .	46
12	Estimates, estimated standard errors and Z-values for conditional variance of $Z_{BMI}$ at different time points in the latent variable model . . . . .	47



## LIST OF FIGURES

1	Plots of smooth terms in a generalized additive model for the smoking cessation study. (The dashed lines indicate plus and minus two pointwise standard deviations.) . . . . .	31
2	Plot of Pearson residuals for the proposed latent variable model versus missing patterns. . . . .	48

## 1.0 INTRODUCTION

Longitudinal studies are increasingly common in public health and medicine, in which, each subject is to be observed at a fixed number of times. Consequently, subjects commonly have missing data due to missed visits. A subject is called a dropout when the response variable is observed through a certain visit and is missing for all subsequent visits (Diggle, Liang, and Zeger 1994[1]). Otherwise, the missing pattern is called arbitrary or intermittent.

In general, the missingness mechanism concerns whether the missingness is related to the study variables or not. Little and Rubin (1987)[2] divide these mechanisms into three classes: Missing Completely at Random (MCAR), Missing at Random (MAR), and Nonignorable (NI). Suppose  $\mathbf{Y}$  is a data matrix that includes observed and missing data and let  $\mathbf{Y}^o$  be the set of observed values of  $\mathbf{Y}$ ,  $\mathbf{Y}^u$  be the set of unobserved or missing values of  $\mathbf{Y}$  and let  $\mathbf{R}$  be the missing data indicator matrix:  $R_{ij} = 1$ , if  $Y_{ij}$  is observed; and  $R_{ij} = 0$ , if  $Y_{ij}$  is missing. Missing Completely At Random (MCAR) indicates that the missingness is unrelated to the values of any variables, whether missing or observed, so:  $Pr(\mathbf{R}|\mathbf{Y}) = Pr(\mathbf{R})$  for all  $\mathbf{Y}$ . Generally one can test whether MCAR conditions can be met by comparing the distribution of the observed data between the observed cases and missing cases (Little 1988[3]). Unfortunately this is hard when there are few cases as there can be a problem with Type I errors. Non-Ignorable (NI) missingness is at the opposite end of the spectrum. In this case, the missingness is related to the missing values. It is nonrandom and is not predictable from any one variable in the data set, that is,  $Pr(\mathbf{R}|\mathbf{Y}) \neq Pr(\mathbf{R})$  for all  $\mathbf{Y}$  and  $Pr(\mathbf{R}|\mathbf{Y})$  depends on  $\mathbf{Y}^u$ . Missing At Random (MAR) is between these two extremes. It requires that the cause of the missing data is unrelated to the missing values, but may be related to the observed values of other variables, that is:  $Pr(\mathbf{R}|\mathbf{Y}) = Pr(\mathbf{R}|\mathbf{Y}^o)$  for all  $\mathbf{Y}^u$ . MAR and MCAR are both said to be ignorable missing data mechanisms.

A simple solution for the modelling of dropout data is ‘last observation carried forward’. As the name indicates, this method replaces the dropouts with the last observed measurement. A refinement of the method would be to estimate a time-trend, either for an individual subject or for a group of subjects allocated to a particular treatment, and to extrapolate not at a constant level, but relative to this estimated trend. Thus, if  $Y_{ij}$  is the last observed measurement on the  $i$ th subject at the  $j$ th time point,  $\hat{\mu}_i(t)$  is the estimated time-trend and  $R_{ij} = Y_{ij} - \hat{\mu}_i(t_j)$ , the method would impute the missing values as  $Y_{ij} = \hat{\mu}_i(t_k) + R_{ij}$  for all  $k > j$  (here  $j$  and  $k$  are time points). The last observation carried forward is mostly used in the pharmaceutical industry, and elsewhere, in the analysis of randomized parallel group trials for which a primary objective is to test the null hypothesis of no difference between treatment groups. But in general, this method is not recommended.

Another very simple way of dealing with missingness is ‘complete case analysis’. Using this approach, one discards all incomplete sequences. When the missingness process is not related to the measurement process, it is obviously wasteful of data. But when these two processes are related, it has the potential to introduce bias because the complete case cannot be assumed to be a random sample with respect to the distribution of the outcomes. So this approach is not recommended as a general method either, except in the case where the interest is focused on the sub-population of completers.

Now we summarize some approaches to parametric modelling of longitudinal data with potentially informative missingness. There mainly exist three methods: selection models, pattern mixture models and random effects models. In a selection model, the joint distribution of  $\mathbf{Y}^o$  and  $\mathbf{R}$  is factored into the marginal distribution of  $\mathbf{Y}^o$  and the conditional distribution of  $\mathbf{R}$ , given  $\mathbf{Y}^o$ , that is,  $Pr(\mathbf{Y}^o, \mathbf{R}) = Pr(\mathbf{Y}^o)Pr(\mathbf{R}|\mathbf{Y}^o)$ . The terminology is due to Heckman (1976)[4], and conveys the notion that dropouts are selected according to their measurement history. Pattern mixture models, introduced by Little (1993)[5], work with the factorization of the joint distribution of  $\mathbf{Y}^o$  and  $\mathbf{R}$  into the marginal distribution of  $\mathbf{R}$  and the conditional distribution of  $\mathbf{Y}^o$  given  $\mathbf{R}$ , that is,  $Pr(\mathbf{Y}^o, \mathbf{R}) = Pr(\mathbf{R})Pr(\mathbf{Y}^o|\mathbf{R})$ . From a theoretical point of view, it is always possible to express a selection model as a pattern mixture model and vice versa, as they are simply alternative factorizations of the same joint distribution. Random effects models are extremely useful in longitudinal data analysis. The

idea under this method is that a subject’s pattern of responses in a study is likely to depend on some unobservable characteristics. These unobservable characteristics are then included in the model as random variables, that is, as random effects. A simple formulation of this kind of model would be to postulate a bivariate random effect,  $\mathbf{U} = (U_1, U_2)$  and to model the joint distribution of  $\mathbf{Y}^o$ ,  $\mathbf{R}$  and  $\mathbf{U}$  as  $f(\mathbf{y}^o, \mathbf{r}, \mathbf{u}) = f_1(\mathbf{y}^o|u_1)f_2(\mathbf{r}|u_2)f_3(\mathbf{u})$ . This assumes that  $\mathbf{Y}^o$  and  $\mathbf{R}$  are conditionally independent given  $\mathbf{U}$ . In terms of Little and Rubin’s hierarchy, the dropouts in the above equation are completely random if  $u_1$  and  $u_2$  are independent, whereas if  $u_1$  and  $u_2$  are dependent then, in general, the dropouts are informative.

The latent class model is another approach for the modelling of missing data and can be framed as a kind of pattern mixture model, in which latent classes connect the observed outcomes and the missing patterns in the likelihood function. Before explaining more about the latent class model, we first introduce the concept of latent variable models. For latent variable models, there are two types of variables to be considered. The variables, which can be directly observed, are called ‘manifest’ variables or responses and those which cannot be observed and represent the constructs of interest, are the ‘latent’ variables. In practice, the dimension of the latent variables is much smaller than that of the manifest variables. There are two assumptions underlying the latent variable models. One is that the values of the manifest variables are the results of an individual’s latent variables. The other is that the manifest variables are independent after controlling for the latent variables. This second assumption is also called ‘local independence’. According to Bartholomew and Knott (1999)[6], there are four kinds of latent variable models based on the distributions of the latent and manifest variables: factor analysis, latent trait analysis, latent profile analysis, and latent class analysis (see Table 1). In some literatures, a model with categorical latent variables is also referred to as a latent class model. To illustrate the local independence in a basic latent class model, we assume that there are categorical response variables  $(Y_1, Y_2, \dots, Y_T)$ , and a latent categorical variable  $Z$  such that for each possible sequence of response outcomes  $(y_1, y_2, \dots, y_T)$  and each category  $z$  of  $Z$ ,  $Pr(Y_1 = y_1, \dots, Y_T = y_T|Z = z) = Pr(Y_1 = y_1|Z = z) = \dots = Pr(Y_T = y_T|Z = z)$ . So a latent class model summarizes probabilities of classification  $Pr(Z = z)$  in the latent classes as well as conditional probabilities  $Pr(Y_t = y_t|Z = z)$  of outcomes for each  $Y_t$  within each

Table 1: Four kinds of latent variable models.

Manifest Variables	Latent Variables	
	Continuous	Categorical
Continuous	Factor Analysis	Latent Profile Analysis
Categorical	Latent Trait Analysis	Latent Class Analysis

latent class. More generally, the latent variable  $\mathbf{Z}$  can be multivariate. Similarly, we can derive the conditional distributions of the manifest variables for the model with continuous latent variables.

Missing data is a common issue encountered in the analysis of longitudinal data. In the behavioral intervention setting, missed visits and/or loss to follow up can be extremely problematic. In this area, missed visits can be assumed to be a result of failure of the intervention, sustained lack of interest in the study or decreased desire to change the behavior. For smoking cessation and weight loss studies, these are common issues that must be dealt with at the data analysis phase. For example, Perkins, et al., 2001[7] conducted a weight concern with smoking study. The purpose of this study was to determine if cognitive-behavioral therapy can reduce weight concerns and increase the success of smoking cessation. The study included 219 women who were randomized to one of three groups: i) behavioral weight control to prevent weight gain (weight control); ii) cognitive-behavioral therapy to reduce concerns (CBT); or iii) nonspecific social support (standard), which involved a discussion of weight. Participants were assessed for smoking abstinence, a binary measure, at 4-weeks postquit and 12-months postquit. However, the outcomes at the second time point for some women were missing due to dropout. The assumption in the smoking cessation literature is that these women were smoking so that all missing outcome values are set equal to zero (0 = smoking; 1 = not smoking) for the purposes of analyses. However this approach can introduce bias because not every woman who drops out is smoking. Since women who smoked again after quitting were more likely to drop out of the study, the dropout may be informative and this is the problem that we want to address in the first part of the dissertaion.

As previously discussed, there are two types of missing patterns: monotone missing data or dropout and non-monotone or intermittent missing data. The second part of the dissertation is focused on developing a latent variable model that allows informative intermittent missingness. The example data come from a study comparing a family-based program with usual care for the treatment of severe pediatric obesity. Originally 172 obese children and a parent or guardian living in the same house as the child were included in the investigation. However, individuals with missing baseline data were not included, so the final sample included 133 children. Subjects were randomized to each of the two groups: treatment group (68 subjects) and usual care group (65 subjects) and then followed for 18 months (the first 6 months were the treatment period). Interest centers on the difference in the children's body mass index (BMI) in these two groups. As in most behavioral weight loss studies, the children who were not successful at losing weight are more likely to miss the assessments, so it is reasonable to consider the data as being subject to informative missingness.

The proposed work will focus on two methods for the analysis of data where the outcome is subject to missingness. First, we propose a latent class model for longitudinal binary response data with informative dropout. The latent variable is used as a mechanism to induce independence between the outcome and the missing status. Thus, in the proposed latent class model, the dropout process and response process are assumed to be independent given a latent class. Because this assumption cannot be verified, we will assess the sensitivity by comparing the proposed model with other models such as the shared parameter model and weighted GEE. In the proposed latent class model, the observed data are divided into two latent classes: a special class in which subjects have deterministic outcomes (in the women's smoking cessation study, we assume that the subjects in the special class are in smoking status) and a second one in which the outcomes can be modeled using logistic regression. Latent class models are similar to pattern mixture models. But the proposed approach is useful especially when the sample size is small and there are a large number of missing patterns though the idea of the proposed latent class model can only be applied to some special data sets.

Secondly, we propose a latent variable model for informative intermittent missingness which is an extension of Roy's (2003)[8] latent dropout class model. In our model, the value

of the latent variable is affected by the missing pattern and it is also used as a covariate in modeling the longitudinal response. Using this approach, the latent variable links the longitudinal response and the missing process. In our model the latent variable is continuous instead of categorical and we assume that it is from a normal distribution with unity variance. To simplify the analysis for intermittent missing patterns, we define two variables: one for the dropout time, and the other for the number of missing time points before dropout. The EM algorithm is used to obtain the estimates of the parameter we are interested in and Gauss-Hermite quadrature is used to approximate the integration of the latent variable (Sammel, et al., 1997[9]).

## 2.0 LITERATURE REVIEW FOR MISSING DATA ANALYSIS AND LATENT VARIABLE MODELS

### 2.1 APPROACHES FOR LONGITUDINAL MISSING DATA

#### 2.1.1 GENERALIZED ESTIMATING EQUATIONS (GEE)

Generalized estimating equations (GEE) are widely used for the analysis of longitudinal data (Liang and Zeger 1986[10]) since they have many advantages over standard approaches. This approach does not require the complete specification of the joint distribution of the repeated responses but rather only the first two moments, making it easier to apply and to extend to outcomes of various types. In GEE, the correlations among the outcomes are treated as nuisance parameters. Correct specification of the variance-covariance structure improves the precision of the estimate. The GEE approach also yields consistent marginal regression parameter estimates when the responses are MCAR because it solves the problem of missing data by simply basing inferences on the observed responses. But when the data is not MCAR, the standard GEE estimates can yield biased regression parameter estimates and hence fail to provide consistent estimates.

Xie and Paik (1997)[11] present an approach for the missing covariate problem in the GEE model when the outcomes are binary and the probability of missingness depends on the observed outcomes and covariates. To deal with the missing covariate problem, the proposed method replaces the missing terms in the estimating functions with consistent estimates obtained from the completely observed units. In this method, it is assumed that the covariates and the random process that causes the covariates to be missing are independent when conditioning on the observed data. Denote  $\mathbf{Y}$ ,  $Z$ , and  $X$  as the outcomes, the subset



of the completely observed covariates and the subset of the partially observed covariates, respectively. Define  $r_i$  to be the observation indicator for  $X_i$ , that is,  $r_i = 1$  if  $X_i$  is observed, and  $r_i = 0$  otherwise, then under the conditional independence assumption,  $f(X|\mathbf{Y}, Z, r = 1) = f(X|\mathbf{Y}, Z, r = 0)$ . Additionally, the outcomes are assumed to be either completely observed or missing completely at random. The estimate of the regression coefficients is shown to be consistent and asymptotically normally distributed.

Another approach developed recently for handling missing data within the GEE framework, is a method based on weighted generalized estimating equations (WGEE). Preisser, Galecki, Lohman and Wagenknecht (2000)[12] propose a WGEE approach for incomplete longitudinal binary outcomes, which are dropouts and MAR. In their method, if the dropout mechanism is specified correctly, the unbiased estimates of parameters in the model for the marginal means can be given by the observed responses. Lipsitz, Molenberghs, Fitzmaurice and Ibrahim (2000)[13] presented a modified GEE for handling missing binary response data. The method is less biased than the standard GEE when data are MAR and the working correlation structure is the true correlation structure. Although, this method provides consistent estimates when the data are MCAR, the estimates of the regression parameters may not be consistent for MAR. The proposed modification uses Gaussian estimates of the correlation parameters, i.e., the estimating function that yields an estimate of the correlation parameters is obtained from the multivariate normal likelihood.

Yi and Cook (2002)[14] developed inverse probability-weighted second-order estimating equations for monotone missing data arising in clusters. This approach facilitates consistent estimation of the marginal mean parameters and association parameters under specified assumptions. For computational reasons, they consider using a weighted alternating logistic regression algorithm for the association parameters of the response distribution.

### 2.1.2 LIKELIHOOD-BASED METHODS

Likelihood-based methods are the most common procedures for modelling marginal proportions in longitudinal binary outcomes, in which the parameters describing the association of an individual's repeated measures are regarded as nuisance parameters. In a formal sense

there is no difference between maximum likelihood for incomplete data and maximum likelihood for complete data. As defined in the introduction,  $\mathbf{Y} = (\mathbf{Y}^o, \mathbf{Y}^u)$ , and  $\mathbf{R}$  is the missing-data indicator matrix that identifies the pattern of missing data. We also define  $\theta$  and  $\psi$  to be the vectors of parameters for the densities of  $\mathbf{Y}$  and  $\mathbf{R}$  separately. Then the full likelihood is a function of  $\theta$  and  $\psi$  proportional to  $f(\mathbf{y}^o, \mathbf{r}|\theta, \psi)$ :

$$L_{full}(\theta, \psi) \propto f(\mathbf{y}^o, \mathbf{r}|\theta, \psi), \quad (2.1)$$

where  $f(\mathbf{y}^o, \mathbf{r}|\theta, \psi)$  is obtained by integrating  $\mathbf{Y}^u$  out of the density  $f(\mathbf{y}, \mathbf{r}|\theta, \psi)$ . Maximum likelihood estimates are obtained by maximizing (2.1), and a large sample covariance matrix for the parameters can be estimated using the inverse of the information matrix obtained by differentiating the log-likelihood twice with respect to  $(\theta, \psi)$  or using the bootstrap.

**2.1.2.1 SHARED PARAMETER MODELS** One approach for dealing with informative missing data is the shared parameter model. Various authors have proposed shared random effect models for longitudinal data subject to informative missing data. Heckman (1979) [15], and Wu and Carroll (1988)[16] developed models for a Gaussian primary response. In these models, the primary response and missingness are modeled separately, and both models are linked by a common random parameter. Such models relaxed the common assumption that the missing data are missing at random. But these models also require other assumptions, and can even lead to a wrong conclusion if these other assumptions cannot be met. Follmann and Wu (1995)[17] proposed an approximate generalized linear model with random effects. Their method can be applied to a variety of distributions for the primary and missing data. The generalized linear model for the primary response is conditioned on the random parameter. The approximation of the generalized linear model is obtained by conditioning on the data that describes the missingness. The assumption of the method is that both the means of the primary response and the variable describing the missingness can be written as a linear function of fixed and random effects. This method approximates a mixed generalized linear model with possibly heterogeneous random effects.

Ten Have, Kunselman, Pulkstenis and Landis (1998)[18] formally proposed logistic regression models for observed longitudinal and missing response components with common

random effect parameters. In the shared parameter models, an important assumption is that the drop-out process and longitudinal outcome process are independent by conditioning on the random effects structure underlying both processes. Actually the random effect parameter in the paper is also a latent variable. In this paper, they conduct a sensitivity analysis by comparing the models with an approximate conditional logit model (Follmann and Wu 1995)[17] and the naive mixed effects logit model. They found that the approximate conditional model does poorly with respect to the between-cluster effect and that the naive model does worse for the within-cluster effect when compared to the shared parameter models.

Albert, Follmann, Wang and Suh (2002)[19] extended the work of Follmann and Wu (1995)[17] and Ten Have et al. (1998). They present a model for longitudinal binary data subject to informative missingness in which a Gaussian autoregressive process rather than a random effect is shared between the response and missing-data mechanism. The paper shows that incorporating within-subject autocorrelation through a latent autoregressive process allows for a richer correlation structure for the repeated binary responses and allows for a more realistic link between the response and missing-data mechanism.

**2.1.2.2 TRANSITION MODELS** Transition models are often used in longitudinal data analysis when the interest is in prediction (Diggle and Kenward 1994)[20]. Cox (1970) [21], and Zeger and Qaqish (1988)[22] have proposed models for characterizing the transition patterns in repeated binary data. Stasny (1987)[23] and Conaway (1993)[24] developed first-order Markov chain models for categorical responses in the presence of nonignorable missing data. Cole, Lee, Whitmore and Zaslavsky (1995)[25] developed an empirical Bayes model for Markov-dependent binary sequences with randomly missing observations. For most cases, the transition probabilities are not the same for every individual. By assuming that the individual transition probabilities are from a common distribution, empirical Bayes models can be used to obtain estimates of the transition probabilities. In the proposed method, the transition probabilities are drawn from a common, new family of bivariate beta prior distributions. Liu, Watermaux and Petkova (1999)[26] proposed likelihood-based methods for analyzing longitudinal binary data with noninformative and informative drop-out. They use a first-order transition model for the outcome and different logit models for the drop-out

process which are functions of the response variable. But this model does not allow for intermittent missing observations. Deltour, Richardson, and Le Hesran (1999)[27] proposed stochastic algorithms for approximate maximum likelihood estimation for Markov models with intermittent missing data. Albert (2000)[28] extended their method by developing a transitional model that allows for (i) a more flexible transitional model with  $k$ th order Markov dependence, (ii) both dropout and intermittent nonignorable missingness, and (iii) longitudinal binary data sets with a large number of observations per subject. They also propose an EM algorithm for parameter estimation.

**2.1.2.3 OTHER METHODS** Fitzmaurice, Laird and Zahner (1996)[29] proposed multivariate logistic models for binary responses with dropouts. They assume that nonresponse depends on covariates and on both the observed responses and the value of the unobserved response. The association between the binary responses is modeled in terms of conditional log odds ratios. They also introduced some simple procedures for identifying nonignorable models when the response variable is discrete. In 1999, Lipsitz, Ibrahim and Fitzmaurice[30] considered likelihood methods when the outcome observed over time is binary and the covariates of interest are categorical. They assume that the missing data are MAR. Because both the response and covariates are categorical, they obtain the maximum likelihood parameter estimates using the EM algorithm with the weights proposed in Ibrahim (1990)[31]. When the percentage of missing data is low, they consider the parameters of the covariate distribution as nuisance parameters. But when the percentage is high, the estimation of the parameters of interest under this assumption may be unstable. To address this case, they develop a conditional model for the covariate distribution. In these conditional models, changes in the binary responses over time are characterized by the conditional probability of success at time  $t$  given the covariates and the previously observed binary response.

## 2.2 LATENT VARIABLE MODELS

Latent class models are measurement models for categorical variables. The basic assumption of latent class analysis is that the total population can be subdivided into several subgroups (latent classes) that cannot be directly observed. Each individual of the total population belongs to one, and only one, class of a categorical latent variable. Thus, the latent classes are exhaustive and mutually exclusive. Furthermore, it is assumed that the observed variables used to measure the unobserved latent variable are mutually independent given a latent class which is referred to as the assumption of local independence.

Latent class models (Lazarsfeld and Henry 1968[32]; Goodman 1974[33]; McCutcheon 1987[34]) have been used in a wide range of biomedical settings. Lindsay, Clogg and Grego (1991)[35] used a simple latent class model for item analysis to construct mixture models. That is, they model the population as consisting of a finite set of groups, each of which is homogeneous. The conventional latent class model can be described as follows: for each of  $N$  subjects, indexed by  $i$ , we make  $J$  dichotomous (0-1) measurements, say  $(y_{i1}, y_{i2}, \dots, y_{iJ})^T = \mathbf{y}_i$ . Assume that the subjects are drawn from a population with  $v$  latent classes, each class consisting of homogeneous individuals. The theoretical proportions of being in each latent class are  $\pi_1, \dots, \pi_v$ , with  $\sum_t \pi_t = 1$  and  $0 < \pi_t < 1$ . Conditioning on a subject being in the  $t$ th latent class, let the probability of the response vector,  $\mathbf{y}_i$ , be  $q_t(\mathbf{y}_i) = P[\mathbf{Y}_i = \mathbf{y}_i | t] = \prod_{j=1}^J (\lambda_{j|t})^{y_{ij}} (1 - \lambda_{j|t})^{(1-y_{ij})}$ . That is, within each latent class, the probabilities for the vector of responses are independent with unknown and item-specific success probabilities,  $\lambda_{j|t}$ . If the probability of being in latent class  $t$  is  $\pi_t$ , the overall likelihood is  $L(\lambda, \pi) = \prod_i \sum_t \pi_t q_t(\mathbf{y}_i)$ .

Hadgu and Qu (1998)[36] first applied a latent class modelling approach incorporating random effects and covariates to diagnostic testing for sexually transmitted diseases. Suppose the latent classes are denoted by  $\delta = 1$  or  $\delta = 0$  for the presence or absence of disease, respectively. In a latent class analysis with random effects, they introduce a latent variable,  $t$ , which summarizes the attributes of the subject or the diagnostic test that are not explained by the disease status alone. Assume that  $t$  is distributed according to a standard normal distribution. For the  $i$ th diagnostic test, a test result is denoted by  $Y_i$ . The relationship between the outcome and latent variable can be expressed as:  $Pr(Y_i = 1 | \delta, b) = \Phi(a_{i\delta} + b_{\delta}t)$ ,

where  $\Phi$  is the cumulative density function of the standard normal variate and  $a_{i\delta}$  and  $b_\delta$  are two parameters. When  $b_\delta = 0$ , this model reduces to the traditional latent class model. In this setting, the probability function is given by  $Pr(\mathbf{Y}|\delta) = \int_{-\infty}^{\infty} \prod_{i=1}^p \Phi(a_{i\delta} + b_\delta t)^{Y_i} [1 - \Phi(a_{i\delta} + b_\delta t)]^{1-Y_i} \phi(t) dt$ . For a latent class model with random effects and covariates, the probability of a positive response given the true disease status  $\delta$ , random effect  $t$ , the  $i$ th diagnostic test, and covariates  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})'$ , ( $m$  denotes the number of covariates), can be expressed as  $Pr(Y_i = 1|\delta, t, \mathbf{x}_i) = \Phi(a_{i\delta} + b_\delta + \mathbf{c}'_\delta \mathbf{x}_i)$ , where  $\mathbf{c}_\delta$  is an  $m \times 1$  vector of coefficients.

Roeder, Lynch and Nagin (1999)[37] developed a technique for handling uncertainty in latent class assignment by building a complex mixture model for the full dataset. The methods that they develop provide an extension compared with other papers in two ways: they allow for the uncertainty of latent class membership; and they develop a model for multivariate analysis of risk factors. In practice, there are two stages for a mixture model approach: in the first stage, response variables are used to categorize individuals by the latent trait; then standard methods of analysis are used to identify cross-group differences. Clogg (1995)[38] pointed out that ignoring the uncertainty of latent trait assignment would result in inherent dangers in the classify-analyze paradigm.

Because latent classes cannot be observed directly, problems arise when it is not clear how many classes are appropriate. Previous work has shown that the Pearson  $\chi^2$  statistic and the log likelihood ratio  $G^2$  statistic are not valid test statistics for evaluating latent class models. Garrett and Zeger (2000)[39] developed and illustrated graphical methods for choosing an appropriate number of classes in latent class models. They develop procedures for assessing Markov chain Monte Carlo convergence and for selecting the number of categories for the latent variable based on evidence in the data using Markov chain Monte Carlo techniques.

Latent class models have been applied widely in the medical area for diagnoses. In 2002, Reboussin, Miller and Lohman[40] applied latent class models for missing data. The data are multiple longitudinal binary health outcomes with multiple-cause non-response when the data are missing at random. They apply the latent transition model of Reboussin et al. (1998[41], 1999[42]) that models the probability of being in a current stage conditionally on the prior stage and covariates using a baseline category logistic regression model. They

extend the estimating equations approach of Robins and co-workers (1995)[43] to latent class models by reweighting the multiple binary longitudinal outcomes by the inverse probability of being observed. This results in consistent parameter estimates when the probability of non-response depends on observed outcomes and covariates (missing at random) assuming that the model for non-response is correctly specified. Robust variance estimates are derived which account for the use of a possibly misspecified covariance matrix, estimation of missing data weights, and estimation of latent class measurement parameters. In this paper, they discuss the issue that latent class models are not verifiable which is similar to random-effects models, and the paper also ignores the uncertainty of latent class membership. From an application standpoint one should pay attention to the fact that while the introduction of additional classes may result in a statistical improvement in a fit, the classes themselves may be clinically uninformative.

Lin, Turnbull, McCulloch and Slate (2002)[44] applied latent class models to a censored survival outcome. The proposed model easily accommodates highly unbalanced longitudinal data and recurrent events. There are two levels of structure in the latent class joint model. First, the uncertainty of latent class membership is specified through a multinomial logit model. Second, the class-specific trajectory and event process are specified parametrically and semiparametrically, under the assumption of conditional independence given the latent class membership. In the article, they provide empirical methods to check this conditional independence assumption. They use a likelihood approach to obtain parameter estimates via the EM algorithm. Patterson, Dayton and Graubard (2002)[45] use jackknife as a method of estimating standard errors for the latent class model parameters.

A latent dropout class model is proposed by Roy (2003)[8] for modeling longitudinal data with nonignorable dropouts. Pattern mixture models are very useful for nonignorable missing data, but are not feasible when there are too many dropout patterns and the sample size is not large, which leads to some patterns with very few subjects. The ideas of latent dropout class models are based on the assumption that a small number of latent classes exist behind the sparse observed dropout times and that the probability of being in a given class is determined by the time of dropout. Therefore, the likelihood for the response is a mixture of the latent dropout classes, instead of over the observed dropout times themselves as is the

case for the pattern mixture model. Parameter estimates are obtained using the method of maximum likelihood and a modified Newton-Raphson algorithm is proposed for it.



### **3.0 A LATENT CLASS MODEL FOR LONGITUDINAL BINARY RESPONSES WITH INFORMATIVE DROPOUT**

Nonignorable missing data is a common problem in longitudinal studies. Latent class models are attractive for simplifying the modeling of missing data when the data are subject to either a monotone or intermittent missing data pattern. In our study, we propose a new latent class model for categorical data, dividing the observed data into two latent classes; one class in which the outcomes are deterministic and a second one in which the outcomes can be modeled using logistic regression. In our model, the latent classes connect the longitudinal responses and the missingness process under the assumption of conditional independence. Parameters are estimated by the method of maximum likelihood based on the above assumption and tetrachoric correlation (le Cessie 1994[46]) between responses. We compare the proposed method with a shared parameter model and weighted GEE using both a clinical trial data set and simulations. The results show that our method and the shared parameter model are similar and better than the weighted GEE model. Although the results obtained using the proposed method and the shared parameter model are similar, our proposed method is simpler to implement and can also be used for intermittent missing data.

#### **3.1 INTRODUCTION**

Missing data is a common issue encountered in the analysis of longitudinal data. In the behavioral intervention setting, missed visits and/or loss to follow up can be extremely problematic. In this area missed visits are assumed to be a result of failure of the intervention, sustained lack of interest in the study or decreased desire to change the behavior. For smoking

cessation and weight loss studies, these are common issues that must be dealt with at the data analysis phase. For example, Perkins, et al., 2001[7] conducted a weight concern with smoking study. The purpose of this study is to determine if cognitive-behavioral therapy can reduce weight concerns and increase the success of smoking cessation. The study includes 219 women who were randomized to one of three groups: i) behavioral weight control to prevent weight gain (weight control); ii) cognitive-behavioral therapy to reduce concerns (CBT); or iii) nonspecific social support (standard), which involved a discussion of weight. Participants were assessed for smoking abstinence, a binary measure, at 4-weeks postquit and 12-months postquit. However, the outcomes at the second time point for some women were missing due to dropout. The assumption in the smoking cessation literature is that these women were smoking so that all missing outcome values are set equal to zero (0 = smoking; 1 = not smoking) for the purposes of analyses. However this assumption can introduce bias since not every woman who drops out is smoking. Since women who smoked again after quitting are more likely to drop out of the study, the dropout may be informative and this is the problem that we want to address in this part.

For informative missingness, in which the missing status depends on unknown outcome values, there are two main methods: selection models and pattern mixture models. For selection models, the joint distribution of the outcome and missingness is partitioned into the marginal distribution of the outcome and the conditional distribution of the missingness given the outcomes. As an alternative to selection models, pattern mixture models work with the factorization of the joint distribution of the outcome and missingness into the marginal distribution of missingness and the conditional distribution of the outcome given missingness. Latent class models are another approach for informative missingness, and can be framed as a special case of a pattern mixture model, in which latent classes connect the observed outcomes and the missing patterns in the likelihood function.

For latent variable models, variables are classified as ‘manifest’ when they can be directly observed and as ‘latent’ when they cannot be directly observed. A latent class model is a type of latent variable model with latent variables being categorical. Latent class models have been applied widely in the medical area for diagnoses (e.g. Hadgu and Qu, 1998[36]; Garret et al., 2000[39]). Recently, latent class models have also been used for dealing with

missing data. Reboussin et al. (2002)[40] proposed a latent class approach for multiple binary longitudinal outcomes subject to missing at random. The idea behind this method is to reweight the binary outcomes by the inverse probability of being observed, which is an extension of Robins, et al. (1995)[43]’s estimating equation approach. Roy (2003)[8] proposed a latent dropout class model for continuous data with nonignorable dropouts. The ideas of latent dropout class models are based on the assumption that a small number of latent classes can be used to represent the sparse observed dropout times and that the probability of being in a given class is determined by the time of dropout. Therefore, the likelihood for the response is a mixture of latent dropout classes, rather than over the observed dropout times themselves, as is the case for the pattern mixture model.

We propose a latent class model for longitudinal binary response data with informative dropout. The latent variable is used as a mechanism to induce independence between the outcome and the missing status. Thus in the proposed latent class model, the dropout process and response process are assumed to be independent given a latent class. Because this assumption cannot be verified, we will assess the sensitivity by comparing the proposed model with other models such as the shared parameter model and weighted GEE. In the proposed latent class model, the observed data are divided into two latent classes: a special class in which subjects have deterministic outcomes (in the women’s smoking cessation study, we assume that the subjects in the special class are in smoking status) and a second one in which the outcomes can be modeled using logistic regression. Under these assumptions there is no need to choose an appropriate number of latent classes. As we mentioned before, latent class models are similar to pattern mixture models. But latent class models can handle intermittent missing data in the same way as for monotone missing data. It is very useful especially when the sample size is not large, but there are a large number of missing patterns because it avoids the case of very few subjects in one missing pattern which is a problem in the traditional pattern mixture models.

The proposed latent class model is addressed in Section 3.2. The shared parameter model (Ten Have et al., 1998[18]) and the weighted GEE (Robins et al., 1995[43]) are also described in Section 3.2. In Section 3.3, simulation results are presented for these three models. The results obtained when applying these models to the women’s smoking cessation data are

presented in Section 3.4. A discussion is provided in the last section.

## 3.2 MODELS

We consider bivariate binary outcomes. For  $n$  individuals, each one can be expressed as  $\mathbf{Y}_i = (Y_{i1}, Y_{i2})'$ . We let  $Y_{ij} = 0$  ( $j = 1, 2$ ) denote smoking and  $Y_{ij} = 1$  denote not smoking for  $i$ th subject at time point  $j$ . Because of missingness, some subjects might only have  $Y_{i1}$  or have no outcomes at all. This kind of missing pattern is due to dropout and may be related to unobserved outcomes, resulting in informative drop-outs. Here we consider the settings where  $Y_{i1}$  is always observed and  $Y_{i2}$  could be observed or missing. Let  $R_i$  be the indicator denoting the missing status of subject  $i$ , where  $R_i = 1$  if  $Y_{i2}$  is observed and 0 if  $Y_{i2}$  is missing.

### 3.2.1 Latent Class Model (LCM)

In our latent class model, a latent class is added into the pattern mixture model. We define  $\eta_i$  ( $i = 1, \dots, n$ ) to be  $i$ th subject's latent class, and  $\eta_i = 1$  or 0. For simplicity, we assume that  $\eta_i = 1$ , if subject  $i$  is in a special status, such that  $Pr(\mathbf{Y}_i = (0, 0)' | \eta_i = 1) = 1$ . This means that if subject  $i$  belongs to the class  $\eta_i = 1$ , then the outcomes are  $(0, 0)'$ . In the smoking study, it can be explained as a smoking phenotype, where the subject has more difficulty with smoking cessation when compared with individuals who are not in the same class. Furthermore, when  $\eta_i = 0$  the subject  $i$  is considered to be in regular status. In the following description, we assume the covariate matrix,  $\mathbf{X}$ , is fixed. Thus, any distribution that is mentioned is actually the distribution conditional on  $\mathbf{X}$ . Under these assumptions, the pattern mixture model with the latent class is given by

$$\begin{aligned} Pr(\mathbf{y}, r, \eta) &= \prod_{i=1}^n Pr(\mathbf{y}_i | r_i, \eta_i) Pr(\eta_i | r_i) Pr(r_i) \\ &= \prod_{i=1}^n Pr(\mathbf{y}_i | \eta_i) Pr(\eta_i | r_i) Pr(r_i). \end{aligned} \tag{3.1}$$

Here we let  $Pr(\mathbf{y}_i|r_i, \eta_i) = Pr(\mathbf{y}_i|\eta_i)$ , that is, given the latent class,  $\eta_i$ , the outcome,  $\mathbf{Y}_i$ , is independent of the missingness,  $R_i$ . This is an important assumption which reduces the mathematic complexity for estimation. For the probabilities of latent classes given missingness,  $Pr(\eta_i|r_i)$ , we set  $\mathbf{e} = (e_1, e_0)$ , where  $e_1 = Pr(\eta_i = 0|R_i = 1)$  and  $e_0 = Pr(\eta_i = 0|R_i = 0)$ .

Suppose the conditional outcome probability,  $p_{ij} = Pr(Y_{ij} = 1|\eta_i = 0)$ , can be fit by the following logistic regression,

$$\log \frac{p_{ij}}{1 - p_{ij}} = \beta^T \mathbf{X}_{ij}, \quad (3.2)$$

where,  $i = 1, \dots, n$ ,  $j = 1, 2$ ,  $\mathbf{X}_{ij}$  is a covariate vector of  $Y_{ij}$  and  $\beta$  is a vector of parameters. Note that  $p_{ij}$  depends on the same parameters at different time points (It is easy to extend this so that  $p_{ij}$  depends on different parameters for different  $j$ ). This results in the following model for  $p_{ij}$ ,

$$p_{ij} = \frac{\exp(\beta^T \mathbf{X}_{ij})}{1 + \exp(\beta^T \mathbf{X}_{ij})}. \quad (3.3)$$

We let  $p_{i1}$  and  $p_{i2}$  denote the marginal probabilities at time points 1 and 2 for the  $i$ th subject, and  $s_{imn} = Pr(Y_{i1} = m, Y_{i2} = n)$ ,  $m = 0, 1$ , and  $n = 0, 1$ . To calculate  $s_{i11}$ ,  $s_{i10}$ ,  $s_{i01}$  and  $s_{i00}$  from  $p_{i1}$  and  $p_{i2}$ , we have to consider the correlation between  $Y_{i1}$  and  $Y_{i2}$ . Prentice (1988)[47] developed a method that accounts for the correlation but depends on the marginal probabilities. Here we use the tetrachoric correlation, which is extended from probit marginals (Ashford and Sowden, 1970[48]) to the logistic marginals (le Cessie and van Houwelingen, 1994[46]). The general idea is to obtain  $s_{i11}$  by using bivariate standard normal distributions and tetrachoric series,

$$s_{i11} = p_{i1}p_{i2} + n(g_{i1})n(g_{i2}) \sum_{k=0}^{\infty} \frac{1}{(k+1)!} He_k(g_{i1})He_k(g_{i2})\rho^{k+1}, \quad (3.4)$$

where  $n(u) = (2\pi)^{-1/2} \exp(-u^2/2)$  are the density function of the standard normal distribution,  $\rho$  is the correlation of  $Y_{i1}$  and  $Y_{i2}$ ,  $g_{ij} = \Phi^{-1}(p_{ij})$  with  $\Phi(\cdot)$  being the standard normal cumulative distribution function, and

$$He_k(\nu) = \sum_{i=0}^{\lfloor k/2 \rfloor} \frac{k!}{i!(k-2i)!} (-1)^i 2^{-i} \nu^{k-2i} \quad (3.5)$$

are the Hermite polynomials, where  $[k/2]$  is the largest integer in the range of  $\leq k/2$ . After computing  $s_{i11}$ , we can easily obtain  $s_{i10}$ ,  $s_{i01}$  and  $s_{i00}$  by

$$\begin{aligned} s_{i10} &= p_{i1} - s_{i11}, \\ s_{i01} &= p_{i2} - s_{i11}, \\ s_{i00} &= 1 - p_{i1} - p_{i2} + s_{i11}. \end{aligned} \tag{3.6}$$

Based on the description above, the likelihood for the  $i$ th subject is

$$\begin{aligned} L(\beta, \mathbf{e}, \rho; \mathbf{y}_i, \eta_i, r_i) &= L(\beta, \rho, \mathbf{y}_i | \eta_i) L(\mathbf{e}, \eta | r_i) L(r_i) \\ &= [(Pr(\mathbf{y}_i | \eta_i = 0)e_1 + I(Y_{i1} = 0, Y_{i2} = 0)(1 - e_1))I(R_i = 1) \\ &\quad + (Pr(y_{i1} | \eta_i = 0)e_0 + I(Y_{i1} = 0)(1 - e_0))I(R_i = 0)]L(r_i), \end{aligned} \tag{3.7}$$

where  $I(\cdot)$  is the indicator function, and  $Pr(\mathbf{y}_i | \eta_i = 0)$  and  $Pr(y_{i1} | \eta_i = 0)$  can be obtained from equations (3.3), (3.4) and (3.6). Because the marginal distribution of  $R_i$  does not depend on the parameters we are interested in, it can be ignored when maximizing the likelihood. We use the quasi-Newton method to obtain the estimates of  $\beta$ ,  $\rho$  and  $\mathbf{e}$  by maximizing the marginal likelihood in the equation (3.7). Initial values for estimation may be obtained from PROC GENMOD in SAS by assuming that outcomes are missing completely at random. The standard errors of the estimates are obtained from the inverse of the Hessian matrix of the final marginal likelihood.

The proposed latent class model can also be applied to the intermittent missing data problem resulting in the following likelihood function:

$$\begin{aligned} L(\beta, \mathbf{e}, \rho; \mathbf{y}_i, \eta_i, \mathbf{r}_i) &= [(Pr(\mathbf{y}_i | \eta_i = 0)e_{11} + I(Y_{i1} = 0, Y_{i2} = 0)(1 - e_{11}))I(\mathbf{R}_i = (1, 1)') \\ &\quad + (Pr(y_{i1} | \eta_i = 0)e_{10} + I(Y_{i1} = 0)(1 - e_{10}))I(\mathbf{R}_i = (1, 0)') \\ &\quad + Pr(y_{i2} | \eta_i = 0)e_{01} + I(Y_{i2} = 0)(1 - e_{01}))I(\mathbf{R}_i = (0, 1)')]L(\mathbf{r}_i). \end{aligned} \tag{3.8}$$

Here  $\mathbf{R}_i$  is a  $2 \times 1$  vector,  $(R_{i1}, R_{i2})$ , of indicator variables denoting the missing status of a subject  $i$ , where  $R_{ij} = 1$  if  $Y_{ij}$  is observed and 0 if  $Y_{ij}$  is missing, where  $j = 1, 2$ ; and  $e_{mn} = Pr(\eta_i = 0 | R_{i1} = m, R_{i2} = n)$ ,  $m = 0, 1$  and  $n = 0, 1$ .

### 3.2.2 Shared Parameter Model (SPM)

Ten Have, Kunselman, Pulkstenis and Landis (1998)[18] developed a shared parameter model with a logistic link for longitudinal binary response data to accommodate informative dropout. The model includes two components: observed longitudinal components and dropout components. These two parts share random effects parameters and they are independent after conditioning on the random effects structure. This independence is a critical assumption for this method.

Let  $Z_i$  be the indicator for dropout where  $Z_i$  can take on the values 1 or 2, with  $Z_i = 2$  indicating that a subject does not drop out at all. Let  $\tau_i$  be the random effect vector for the  $i$ th subject, and we let  $\tau_i$  have a multivariate normal distribution with mean 0 and variance-covariance that is an identity matrix with the appropriate dimension.

The resulting marginal likelihood for the  $i$ th subject for both the drop-out component and the longitudinal component is

$$f(\mathbf{y}_i, z_i) = \int f_y(\mathbf{y}_i|\tau) f_z(z_i|\tau) f(\tau) d\tau. \quad (3.9)$$

$Y_{ij}|\tau$  is Bernoulli( $\pi_{ij}$ ), to calculate  $\pi_{ij}$ , then the logistic model is

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \tau_i^T \mathbf{\Sigma} \mathbf{w}_{ij} + \beta^T \mathbf{x}_{ij}, \quad (3.10)$$

where  $\mathbf{x}_{ij}$  and  $\mathbf{w}_{ij}$  are the observed covariate vectors corresponding to the fixed and random effects, respectively, for the  $i$ th subject at time point  $j$ , and  $\mathbf{\Sigma}$  is the Cholesky decomposition of  $\mathbf{\Omega}$ , that is,  $\mathbf{\Omega} = \mathbf{\Sigma}^T \mathbf{\Sigma}$ .

Let  $S_{i1} = 1$  if subject  $i$  drops out between time 1 and 2, and 0 otherwise; and  $S_{i2} = 1$  if the  $i$ th subject does not drop out and 0 otherwise. Further, define:

$$\begin{aligned} \lambda_{ij} &= Pr(Z_i = j | Z_i > j - 1; \tau_i) \\ &= Pr(S_{ij} = 1 | S_{ij'} = 0, j' < j; \tau_i). \end{aligned} \quad (3.11)$$

Let  $\psi_{ij} = \lambda_{ij}(1 - \lambda_{ij})^{-1}$ , where  $\psi_{ij}$  is a continuation ratio, then

$$\begin{aligned} \log[f_z(z_i|\tau)] &= s_{i1} \log \lambda_{i1} + s_{i2} [\log \lambda_{i2} + \log(1 - \lambda_{i1})] \\ &= s_{i1} \log \psi_{i1} + \log(1 - \lambda_{i1})(s_{i1} + s_{i2}) \\ &= s_{i1} \log \psi_{i1} + \log(1 - \lambda_{i1}). \end{aligned} \quad (3.12)$$

It is assumed that  $s_{i1} + s_{i2} = 1$ . In order to obtain  $f_z(z_i|\tau)$ ,  $\log \psi_{ij}$  should be modeled first. Here

$$\log \psi_{ij} = \tau_i^T \Sigma^* \mathbf{w}_{ij} + \rho^T \mathbf{u}_{ij}, \quad (3.13)$$

where  $\Sigma^* = \Sigma + \Delta$ ,  $\Delta$  is an upper triangular matrix, as is  $\Sigma$ , and  $\mathbf{u}_{ij}$  is a vector of covariates specific to the dropout process.

### 3.2.3 Weighted GEE (WGEE)

Robins, Rotnitzky and Zhao (1995)[43]’s weighted GEE approach yields consistent estimates when the responses are MAR. The missing data can be written as

$$\nu_{im_i} = f_m(m_i|\mathbf{y}_i, \mathbf{x}_i, \gamma) = Pr(M_i = m_i|\mathbf{y}_i, \mathbf{x}_i, \gamma), \quad (3.14)$$

where  $M_i = 1 + \sum_{j=1}^2 r_{ij}$ , and  $\gamma$  is the vector of parameters of the nonresponse model. Thus  $\nu_{im_i}$  can be obtained by modelling the dropout mechanism using  $Y_{i1}$  and/or  $\mathbf{X}_i$ .

Suppose  $\mu_i(\beta) = E(\mathbf{Y}_i|\mathbf{X}_i, \beta)$ , then we can partition  $\mathbf{Y}_i$  into the unobserved components  $\mathbf{Y}_i^u$  and the observed components  $\mathbf{Y}_i^o$ . Similarly, we can make the exact same partition of  $\mu_i$  into  $\mu_i^u$  and  $\mu_i^o$ . Then under the weighted GEE approach,

$$\mathbf{U}_\beta(\hat{\beta}) = \sum_{i=1}^n \frac{1}{\nu_{im}} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} [\mathbf{Y}_i^o - \mu_i^o] = \mathbf{0}, \quad (3.15)$$

where  $\mathbf{D}_i = \partial \mu_i^o / \partial \beta$  and  $\mathbf{V}_i$  is a ‘working’ covariance matrix of  $\mathbf{Y}_i^o$ .

## 3.3 SIMULATION RESULTS

We perform a simulation study comparing the proposed method, the shared parameter model and the weighted GEE. We generate data by considering two aspects: the logistic model structure for  $(\mathbf{X}, \mathbf{Y})$  and the missing structure  $(R)$ .

For generating the covariate data, we let  $X_1$  be a standard normal variable, and  $\mathbf{X}_2$  be a bivariate normal variable with mean = (0, 0.2) and variance-covariance given by



$\begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}$ , which will be considered as a time-varying covariate in modeling longitudinal outcomes. For the outcomes, we consider the case of a binary response measured at two time points,  $\mathbf{Y}_i = (Y_{i1}, Y_{i2})'$  with correlation  $\rho$ . To generate data for this scenario, we first generate the continuous variables  $(Z_1, Z_2)'$ , which are from a bivariate standard normal distribution with correlation  $\delta$ . Then we let  $Y_{ij} = 1$  if  $Z_j \leq \Phi^{-1}(p_{ij})$ ;  $Y_{ij} = 0$ , if  $Z_j > \Phi^{-1}(p_{ij})$ , where  $j = 1, 2$ , and  $\Phi(\cdot)$  is the standard normal cumulative distribution function. Here  $p_{ij}$  are the marginal probabilities,  $p_{ij} = Pr(Y_{ij} = 1) = E(Y_{ij})$ , and are obtained from  $logit(p_{ij}) = \beta^T \mathbf{X}_{ij} = \beta_1 t + \beta_2 X_{1i} + \beta_3 X_{1i}^2 + \beta_4 X_{2ij}$ , with  $t = 1$  or  $10$ ;  $\beta_1 = 0.1$ ,  $\beta_2 = 0.2$ ,  $\beta_3 = 0.3$  and  $\beta_4 = 0.4$ . For the purpose of simulation, the correlation,  $\delta$ , is set to be 0.5. According to Emrich, et al. (1991)[49], the correlation between  $Y_{i1}$  and  $Y_{i2}$  is given by,  $\rho = [\Phi(\Phi^{-1}(p_{i1}), \Phi^{-1}(p_{i2}), \delta) - p_{i1}p_{i2}] / [p_{i1}(1 - p_{i1})p_{i2}(1 - p_{i2})]^{1/2}$ .

For the missing structure, we assume a monotone missing data pattern with the binary response at the first time point completely observed. Three missing mechanisms are considered for the response at time point 2: (i) WGEE missing mechanism, in which

$$logit[Pr(R_i = 0|Y_{i1})] = -0.5 - 0.5Y_{i1}.$$

Note that the missingness depends on the observed outcomes so that the missing mechanism under the WGEE is MAR. (ii) SPM missing mechanism. For the SPM mechanism, note that the  $p_{ij} = Pr(Y_{ij} = 1)$  in the  $(\mathbf{X}, \mathbf{Y})$  structure are obtained from

$$logit(p_{ij}) = \beta^T \mathbf{X}_{ij} + \sigma_1 \tau,$$

where  $\tau$  is a normal variable with mean = 0 and variance  $\sigma_1^2 = 1$ . For the missing process,

$$logit[Pr(R_i = 0|X_{i1}, \tau)] = -0.5 - 0.5X_{i1} + \sigma_2 \tau,$$

where  $\sigma_2^2 = \sigma_1^2 + \delta$  and  $\delta = -0.1$ . (iii) LCM missing mechanism, in which we let  $Pr(R_i = 0) = 0.3$ . In the data generation procedure, for WGEE and SPM, we first obtain a full data set, then delete some of  $Y_{i2}$  values according to the missing structure. But in LCM, the procedure is different: after obtaining the missing patterns, we define the latent classes in which we let  $e_1 = Pr(\eta_i = 0|Y_{i2} \text{ observed})$ ,  $e_0 = Pr(\eta_i = 0|Y_{i2} \text{ missing})$  and we assume that

$Y_{i1}$  is always observed. Then we generate the  $(\mathbf{X}, \mathbf{Y})$  structure in each latent class. For the latent classes,  $\eta_i$ ,  $\eta_i = 1$  denotes the case where  $Pr(Y_{ij} = 1|\eta_i = 1) = 0$ , that is, if  $\eta_i = 1$  then  $Y_{ij} = 0$ ; and  $\eta_i = 0$  denotes the standard case where  $Pr(Y_{ij} = 1|\eta_i = 0) = \frac{\exp(\beta^T \mathbf{X}_{ij})}{1 + \exp(\beta^T \mathbf{X}_{ij})}$ . For the simulation, we considered  $e_1 = 0.9$  and  $e_0 = 0.7$  to examine the impact of the heterogeneity of the data on the parameter estimate.

For the simulation study, we consider the three missing mechanisms and sample sizes of 200 with 500 replications (Table 2). The summary measures for a particular parameter are the bias, standard error of mean, square root of mean square error and 95% coverage probability. Table 2 presents the simulation results. With the exception of the SPM, each method performed optimally under its own structure. When the missing mechanism is WGEE, the weighted GEE has the smallest bias, standard error of mean and square root of mean square error while the 95% coverage probabilities are low compared with other methods. When the missing mechanism is SPM, the weighted GEE has low 95% coverage probabilities. But the latent class model has the smallest bias for the ‘time’ variable ( $\beta_1$ ) and the time varying continuous variable ( $\beta_4$ ). When the missing mechanism is LCM, the 95% coverage probabilities are low for all of the parameter estimates under the weighted GEE. The shared parameter model performs well in terms of the mean square error and 95% coverage probability. For the latent class model, its biases are small and the 95% coverage probabilities are large compared with the other models.

In the comparison among the three models, the overall conclusions are that, the weighted GEE has the poorest 95% coverage probability, especially in the LCM missing mechanism for  $\beta_1$  and  $\beta_3$ . The shared parameter model performs well under each of the three missing mechanisms. The proposed latent class model has the most accurate 95% coverage probability, and for data generated by the LCM missing mechanism, its bias is smaller than the other methods. So the latent class model performs well in the greatest number of cases and generally outperforms the shared parameter model and weighted GEE for the informative missingness scenario.

### 3.4 APPLICATION TO THE SMOKING CESSATION STUDY

We illustrate the proposed method – latent class model, the shared parameter model and the weighted GEE previously discussed, using an example from the women’s smoking cessation study (Perkins et al., 2001[7]). This is a longitudinal study designed to assess the effect of weight concern on smoking cessation for women. At enrollment, 219 women met the eligibility criteria. If the women were trying to become pregnant or were following a medically prescribed diet, they were not eligible. All of the participants were randomly divided into three groups: i) behavioral weight control to prevent weight gain (weight control); ii) cognitive-behavioral therapy to reduce weight concerns (CBT); or iii) nonspecific social support (standard), which involved no discussion of weight. Each of the three interventions consisted of ten 90-minute sessions provided over 7 weeks, with two sessions per week during the first 3 weeks and one session per week over the next 4 weeks. Participants were instructed to quit smoking after the fourth session. Follow-up sessions were scheduled at 3, 6, and 12 months postquit for assessment purposes; no treatment was provided in these periods. In this trial, the repeated binary responses of interest are whether the participants are in continuous abstinence or not (1 = yes, 0 = no). Here, continuous abstinence was defined as no relapse since the quit day and relapse was defined as self-report of 7 consecutive days of any smoking at all or an expired-air carbon monoxide (CO) greater than 8 ppm, as widely recommended (Ossip-Klein et al. 1986[50]).

In this study, we focus on the outcomes at two time points, 4-week postquit ( $Y_1$ ) and 12-month postquit ( $Y_2$ ). The 57 women who had missing data at 4-week postquit (also missing at 12-month postquit) were removed from all analyses, leaving 162 subjects (116 subjects have no missing data; 46 subjects were observed at 4-week postquit and missing at 12-month postquit). To identify significant covariates related to outcomes, we carried out a preliminary analysis by using standard GEE (Liang and Zeger, 1986[10]) under the assumption of MCAR. The results showed that the following variables should be included in the models: group (*weight control*, *CBT*) (\*‘standard’ as a control group), time ( $t$ ), age at first cigarette (*age*), change in desire to smoke from prequit to postquit (*desire*), and weight gain in percent (*WtGn*). The latter two variables were time-varying, and both the baseline

(4-week postquit) values and the change from baseline are included for the longitudinal effects. We also fit a generalized additive model (GAM) for the outcomes,  $Y_1$ , with three covariates, ‘age at first cigarette’, ‘change in desire to smoke’ and ‘weight gain in percent’. Figure 1 shows plots of smooth terms in GAM. From the results of GAM, we added square terms for ‘age at first cigarette’ and ‘change in desire to smoke’, in addition to the linear terms.

A summary of outcomes is in Table 3. It shows that missingness at 12-month follow-up is 28.40%. The abstinent rate for subjects without missingness at 12-month follow-up is  $53/(53+63) = 45.68\%$ , which is much less than the abstinent rate (72.84%) at the 4-week postquit. It also shows that the dropout might be related to the unobserved second-time outcomes. Based on these results it is reasonable to consider non-ignorable missingness.

In Table 4, we present the parameter estimates, standard errors and Z-values calculated from them for these three models. These parameter estimates are common fixed effects under these models. The results of these analyses suggest that the latent class model and the shared parameter model have similar results. From them, we can see that the ‘CBT’ group has a larger abstinent rate compared with the ‘standard’ group. ‘Time’ is significant in the models with a decreasing abstinent rate over time. Both of them also show that ‘change in desire to smoke’ is significant with a negative linear parameter estimate and positive square term. From Figure 1, we can see that most values of ‘change in desire to smoke’ are negative or around zero, so increasing ‘change in desire to smoke’ will lead to a lower abstinence rate. Both of the analyses show that there is a linear increase (on the logit scale) in the abstinence rate as the ‘weight gain in percent’ increases. The results obtained from the weighted GEE show that ‘age at first cigarette’ is significant in both linear and square terms, while these factors are not significant in the other models. In the latent class model, we obtain the correlation within the subject,  $\rho = 0.28$  (SE = 0.32, Z = 0.89). Table 5 gives the estimates for  $e_1$  and  $e_0$ . It shows that subjects with missing outcomes have a greater probability ( $1 - 0.93 = 0.07$ ) of being in the special status, that is ‘smoking’ phenotype, than the subjects without missing outcomes ( $1 - 0.96 = 0.04$ ).

### 3.5 DISCUSSION

In this part a latent class model is proposed for the analysis of binary repeated measures outcomes subject to informative dropout. The latent variable is used to induce conditional independence between the outcome and missing status so that standard likelihood techniques can be used to derive the estimators. While this model can be considered as a type of pattern mixture model, the latent class model can fit any type of missing data, monotone or intermittent missingness and work well in the small sample setting. The simulation results provide further support for the use of this method when compared to the shared parameter and weighted GEE models. The results indicate that the proposed model generally has a smaller bias when compared to the shared parameter model; and the coverage probabilities of the latent class variable model are significantly better than those of either of the other two methods.

Each of the three methods can be implemented using a standard statistical package such as S-Plus or R; however, the weighted GEE only requires the input of the weights in the general routine for GEE. Additionally, the shared parameter model requires more computational time than the other two methods. The proposed latent class model requires less computational time as there is no need for integration. Both the latent class and shared parameter models are based on likelihood theory so that likelihood ratio and score tests can be computed.

Roy (2003)[8] proposed a latent dropout class model for continuous responses with non-ignorable dropouts. In our latent class model, we develop a latent class model for categorical responses with nonignorable dropouts. The dropout time is related to the latent class, whose probability is estimated by the MLE. For the relationship within subjects, we use the tetrachoric correlation (le Cessie and van Houwelingen, 1994[46]) for the estimation. Here we focused on monotone missingness, but the method can also be used for intermittent missing data in the same way.

Table 2: Simulation results for weighted GEE (WGEE), shared parameter model (SPM) and latent class model (LCM) (sample size = 200, informative dropout).

Bias SE of Mean $\sqrt{MSE}$ 95% CP	WGEE missing mechanism			SPM missing mechanism			LCM missing mechanism		
	WGEE	SPM	LCM	WGEE	SPM	LCM	WGEE	SPM	LCM
$\beta_1 = 0.1$	$1.1 \times 10^{-3}$	$2.2 \times 10^{-2}$	$5.1 \times 10^{-3}$	$-1.2 \times 10^{-2}$	$2.6 \times 10^{-2}$	$-8.5 \times 10^{-4}$	$-2.7 \times 10^{-2}$	$1.1 \times 10^{-2}$	$4.6 \times 10^{-3}$
	$2.7 \times 10^{-3}$	$2.9 \times 10^{-3}$	$3.1 \times 10^{-3}$	$2.5 \times 10^{-3}$	$3.3 \times 10^{-3}$	$3.5 \times 10^{-3}$	$2.2 \times 10^{-3}$	$3.4 \times 10^{-3}$	$4.1 \times 10^{-3}$
	$2.4 \times 10^{-2}$	$3.4 \times 10^{-2}$	$2.8 \times 10^{-2}$	$2.6 \times 10^{-2}$	$3.9 \times 10^{-2}$	$3.1 \times 10^{-2}$	$3.4 \times 10^{-2}$	$3.2 \times 10^{-2}$	$3.7 \times 10^{-2}$
	0.932	0.970	0.986	0.890	0.952	0.971	0.646	0.978	0.947
$\beta_2 = 0.2$	$1.0 \times 10^{-2}$	$3.5 \times 10^{-2}$	$3.0 \times 10^{-2}$	$-7.2 \times 10^{-3}$	$9.4 \times 10^{-3}$	$-2.3 \times 10^{-2}$	$-4.7 \times 10^{-2}$	$-2.8 \times 10^{-2}$	$6.6 \times 10^{-3}$
	$1.9 \times 10^{-2}$	$2.1 \times 10^{-2}$	$2.1 \times 10^{-2}$	$1.9 \times 10^{-2}$	$2.3 \times 10^{-2}$	$2.0 \times 10^{-2}$	$1.7 \times 10^{-2}$	$2.0 \times 10^{-2}$	$2.2 \times 10^{-2}$
	$1.7 \times 10^{-1}$	$1.9 \times 10^{-1}$	$1.9 \times 10^{-1}$	$1.7 \times 10^{-1}$	$2.0 \times 10^{-1}$	$1.8 \times 10^{-1}$	$1.6 \times 10^{-1}$	$1.8 \times 10^{-1}$	$2.0 \times 10^{-1}$
	0.940	0.937	0.963	0.934	0.956	0.955	0.924	0.994	0.974
$\beta_3 = 0.3$	$6.6 \times 10^{-2}$	$7.4 \times 10^{-2}$	$1.2 \times 10^{-1}$	$-2.6 \times 10^{-2}$	$4.7 \times 10^{-2}$	$3.9 \times 10^{-2}$	$-1.4 \times 10^{-1}$	$-1.3 \times 10^{-1}$	$2.9 \times 10^{-2}$
	$1.3 \times 10^{-2}$	$1.5 \times 10^{-2}$	$1.7 \times 10^{-2}$	$1.2 \times 10^{-2}$	$1.5 \times 10^{-2}$	$1.6 \times 10^{-2}$	$1.1 \times 10^{-2}$	$1.1 \times 10^{-2}$	$1.8 \times 10^{-2}$
	$1.4 \times 10^{-1}$	$1.5 \times 10^{-1}$	$1.9 \times 10^{-1}$	$1.1 \times 10^{-1}$	$1.4 \times 10^{-1}$	$1.5 \times 10^{-1}$	$1.7 \times 10^{-1}$	$1.6 \times 10^{-1}$	$1.7 \times 10^{-1}$
	0.934	0.956	0.951	0.920	0.974	0.982	0.656	0.962	0.970
$\beta_4 = 0.4$	$7.4 \times 10^{-3}$	$9.6 \times 10^{-2}$	$3.5 \times 10^{-2}$	$-5.2 \times 10^{-2}$	$4.6 \times 10^{-2}$	$-3.8 \times 10^{-2}$	$-8.0 \times 10^{-2}$	$4.9 \times 10^{-2}$	$2.0 \times 10^{-2}$
	$1.5 \times 10^{-2}$	$1.7 \times 10^{-2}$	$1.5 \times 10^{-2}$	$1.4 \times 10^{-2}$	$2.0 \times 10^{-2}$	$1.6 \times 10^{-2}$	$1.3 \times 10^{-2}$	$1.7 \times 10^{-2}$	$1.8 \times 10^{-2}$
	$1.4 \times 10^{-1}$	$1.8 \times 10^{-1}$	$1.4 \times 10^{-1}$	$1.3 \times 10^{-1}$	$1.8 \times 10^{-1}$	$1.5 \times 10^{-1}$	$1.4 \times 10^{-1}$	$1.6 \times 10^{-1}$	$1.6 \times 10^{-1}$
	0.932	0.962	0.986	0.932	0.972	0.949	0.888	0.988	0.958

Table 3: Summary of outcomes for the smoking cessation study.

		12-month postquit			
		$y_2$ is missing	$y_2 = 0$	$y_2 = 1$	Total
4-week postquit	$y_1 = 0$	13	26	5	44(27.16%)
	$y_1 = 1$	33	37	48	118(72.84%)
Total		46(28.40%)	63(38.89%)	53(32.72%)	162

Table 4: Marginal parameter estimates, estimated standard errors and Z-values for the smoking cessation study (Modelling  $\Pr(\text{abstinent})$ ).

Par.	WGEE			SPM			LCM		
	Est.	SE	Z	Est.	SE	Z	Est.	SE	Z
$\beta_0$	1.37	1.68	0.81	1.55	2.91	0.53	1.51	3.03	0.50
$\beta_{WC}$	0.77	0.42	1.85	0.69	0.44	1.59	0.71	0.50	1.42
$\beta_{CBT}$	1.07	0.46	2.36	1.08	0.46	2.36	1.01	0.52	1.95
$\beta_t$	-0.21	0.04	-5.73	-0.20	0.03	-5.81	-0.25	0.04	-6.16
$\beta_{age}$	-0.49	0.24	-2.02	-0.46	0.40	-1.14	-0.49	0.42	-1.17
$\beta_{(age)^2}$	0.02	0.009	2.64	0.02	0.01	1.60	0.02	0.01	1.65
$\beta_{desire}$	-0.02	0.007	-3.20	-0.03	0.01	-2.71	-0.03	0.01	-2.79
$\beta_{(dsr)^2}$	0.0007	0.0002	3.50	0.0006	0.0002	3.00	0.0008	0.0002	3.94
$\beta_{WtGn}$	0.24	0.05	5.11	0.28	0.06	5.06	0.28	0.06	4.74

Table 5: Estimates, estimated standard errors and Z-values for the latent classes,  $e_1$  and  $e_0$  under the proposed method for the smoking cessation study.

	Estimate	SE	Z-value
$e_1$	0.96	0.03	28.81
$e_0$	0.93	0.07	12.86

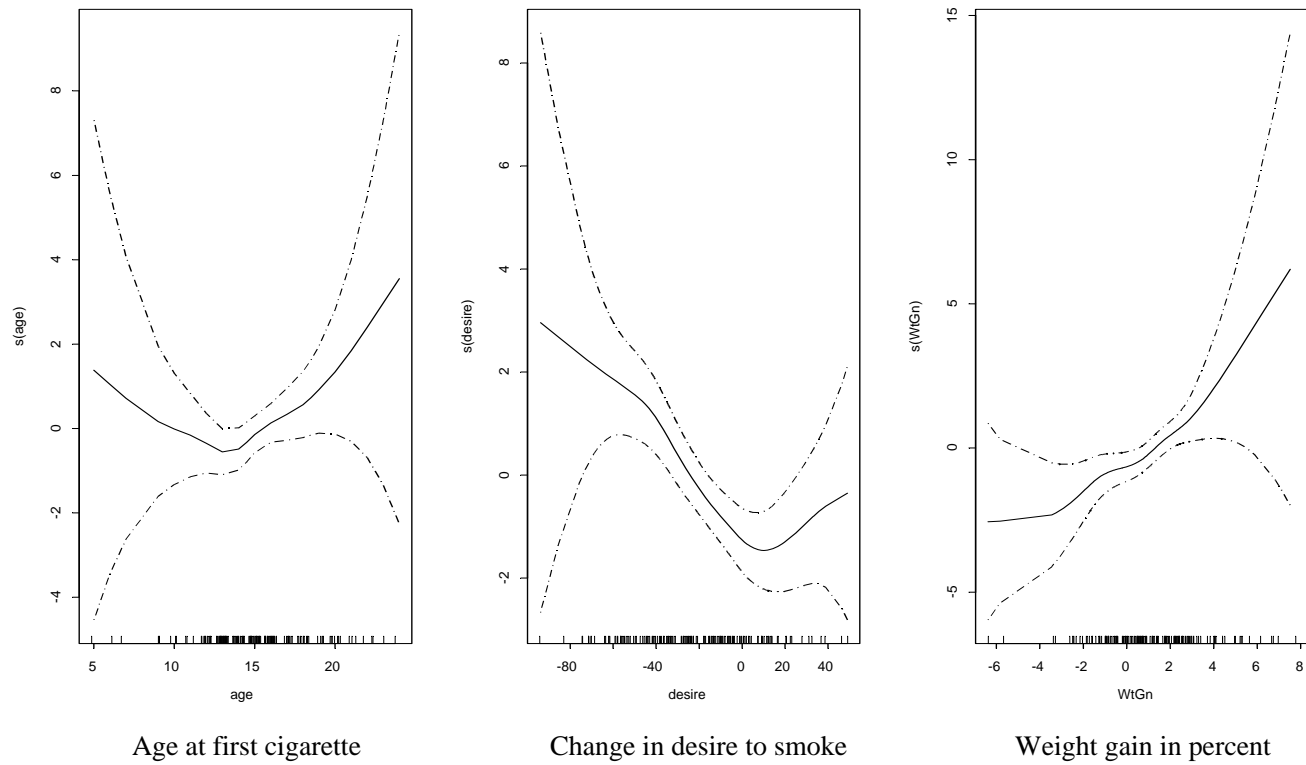


Figure 1: Plots of smooth terms in a generalized additive model for the smoking cessation study. (The dashed lines indicate plus and minus two pointwise standard deviations.)



## 4.0 AN EXTENSION OF LATENT VARIABLE MODEL FOR INFORMATIVE INTERMITTENT MISSING DATA

We propose a latent variable model for informative intermittent missingness in longitudinal studies which is an extension of Roy's (2003) [8] latent dropout class model. In our model, the value of the latent variable is affected by the missing pattern and it is also used as a covariate in modeling the longitudinal response. So the latent variable links the longitudinal response and the missing process. In our model the latent variable is continuous instead of categorical and we assume that it is from a normal distribution with unity variance. The EM algorithm is used to obtain the estimates of the parameter we are interested in and Gauss-Hermite quadrature is used to approximate the integration of the latent variable (Sammel, et al., 1997[9]). The standard errors of the parameter estimates can be obtained from the bootstrap method or from the inverse of the Fisher information matrix of the final marginal likelihood. Comparisons are made to the pattern mixture model in terms of a clinical trial dataset, which is a pediatric obesity study evaluating the effectiveness of a family-based intervention. The proposed method is also compared to generalized estimating equations (GEE). We use the generalized Pearson residuals to assess the fit of the proposed latent variable model.

### 4.1 INTRODUCTION

In longitudinal studies, subjects are followed over time, so missing data are a frequent problem. There are two basic types of missing patterns; one is monotone missing data or dropout in which a subject's data are observed only through a certain time and are missing thereafter

and the other is non-monotone or intermittent missing data in which a subject may return after one or more missed visits. If the missing process depends on unobserved outcomes, according to Rubin (1976)[51] such a missing data process is ‘nonignorable’ or ‘informative’ because likelihood inference is biased if we ignore the missing process. In this paper we developed a latent variable model that allows informative intermittent missingness. The example data come from a study comparing a family-based program with usual care for the treatment of severe pediatric obesity. Originally 172 obese children and a parent or guardian living in the same house as the child were included in the investigation, but we do not consider those who were missing at baseline, so that 133 children are included in the models. Subjects were randomized to each of two groups: treatment group (68 subjects) and usual care group (65 subjects) and then followed for 18 months (the first 6 months were the treatment period). Interest centers on the difference of the children’s body mass index (BMI) in these two groups. As in most behavioral weight loss studies, the children who are not successful at losing weight are more likely to miss the assessments, so it is reasonable to consider the data as being subject to informative missingness.

To account for informative missingness, a number of model-based approaches have been proposed to jointly model the longitudinal outcome and the missing mechanism. Among these approaches, pattern-mixture models (Little, 1993[5]), which factor the joint distribution as the marginal distribution of the mechanism multiplied by the conditional distribution of the response given the mechanism, are commonly used because of their robustness in modeling the missing mechanism. Random effects models for informative missingness have also been investigated by a number of researchers. Ten Have, et al. (1998)[18] presented a shared parameter model for longitudinal binary response data with informative dropout. The model consists of observed longitudinal and missing response components that share random effects parameters. Guo, et al. (2004)[52] proposed a modified pattern mixture model for longitudinal data with dropouts. In their model, the pattern-specific parameters are modeled as random and, conditional on them, the longitudinal outcomes and the dropout process are assumed to be independent.

The other method developed by researchers for informative missing data is latent variable models. Since Goodman’s (1974)[33] development of maximum likelihood procedures

for fitting latent variable models, they have been used for a wide variety of applications. For instance, Clogg (1979)[53] used latent variables to assess agreement among subjects' responses to several survey items; Rindskopf and Rindskopf (1986)[54] used them to analyze characteristics for diagnostic tests; Sammel, et al. (1997)[9] proposed a latent variable model for mixed discrete and continuous outcomes, in which the latent variable accounts for the relationship between different outcomes. In recent years, some authors have used latent variable models for missing data. Moustaki and Knott (2000)[55] discussed the calculation of response propensities by using latent variable models with or without covariates, which is used to weight item responders to account for item non-response when missing data cannot be ignored. Lin, et al. (2004)[56] proposed a latent pattern mixture model, where the mixture patterns are formed from latent classes that link the longitudinal response and the missingness process. They also propose a noniterative approach to assess the assumption of the conditional independence between the longitudinal outcomes and the missingness process given the latent classes. A latent dropout class model is proposed in Roy's (2003)[8] paper. In his paper, it is assumed that there exist a small number of dropout classes. Class membership is unobserved, but the probability of being in a particular latent dropout class is determined by the dropout times. Therefore, the likelihood for the response is a mixture of latent dropout classes, not the observed dropout times themselves as is the case for traditional pattern mixture models. An important assumption of latent variable models for missing data analysis is that conditional on the latent variables, the longitudinal outcome process and missingness process are independent.

We propose a latent variable model for informative intermittent missingness which is an extension of Roy's (2003)[8] latent dropout class model. In our model, the value of the latent variable is affected by the missing pattern and it is also a covariate in modeling the longitudinal response. Using this approach, the latent variable links the longitudinal response and the missing process. In our model the latent variable is continuous instead of categorical and we assume that it is from a normal distribution with unity variance. To simplify the analysis for intermittent missing patterns, we define two variables: one for the dropout time, and the other for the number of missing time points before dropout. The EM algorithm is used to obtain the estimates of the parameter we are interested in and Gauss-

Hermite quadrature is used to approximate the integral of the latent variable (Sammel, et al., 1997[9]).

We describe the proposed latent variable model and the parameter estimation in Section 4.2. In Section 4.3 we apply the proposed model to the children’s obesity data and compare it with the pattern mixture model and GEE, and discuss the assessment of fit of the model. A discussion is provided in the last section.

## 4.2 MODEL SPECIFICATION AND ESTIMATION

### 4.2.1 MODEL SPECIFICATION

Suppose in a longitudinal study with  $K$  repeated measurements and  $N$  individuals,  $Y_{ij}$  denotes the observed vector of responses for the  $i$ th subject observed at the  $j$ th time point. For some reasons, not all subjects have all  $K$  measurements. When this occurs as a result of dropout, the response  $Y_{ij}$  for subject  $i$ , is only observed at time points  $j = 1, \dots, k_i$ ; where  $k_i \leq K$ . But if the data are subject to intermittent missingness, before time point  $k_i$ , there may be additional missing measurements. We use a missing indicator for each of the  $K$  measurements with 1 denoting missing and 0 denoting observed. To simplify the missing status, we define  $\mathbf{R}_i = (R_{i1}, R_{i2})'$ , to be a vector denoting the missing status of subject  $i$ , where  $R_{i1}$  = time point after which the  $i$ th subject drops out and  $R_{i2}$  = number of missing measurements before dropout. The important assumption of the latent variable model is that the longitudinal outcomes and missing process are independent when conditioned on the unobserved latent variable. Let  $b_i$  be the unobserved latent variable for subject  $i$  and  $\mathbf{Y}_i^c = (\mathbf{Y}_i, \mathbf{Y}_i^m)$  denote the complete response vector, which includes the observed response vector,  $\mathbf{Y}_i$ , and the unobserved one,  $\mathbf{Y}_i^m$ . The marginal likelihood for the  $i$ th subject for

both the response and missing components is then

$$\begin{aligned}
f(\mathbf{Y}_i, \mathbf{R}_i) &= \int \int f_y(\mathbf{Y}_i^c | b_i, \mathbf{R}_i) f_b(b_i | \mathbf{R}_i) f(\mathbf{R}_i) dY_i^m db_i \\
&= f(\mathbf{R}_i) \int f_y(\mathbf{Y}_i | b_i) f_b(b_i | \mathbf{R}_i) db_i \\
&\propto \int f_y(\mathbf{Y}_i | b_i) f_b(b_i | \mathbf{R}_i) db_i,
\end{aligned} \tag{4.1}$$

where  $f_y(\mathbf{Y}_i | b_i)$  is the conditional distribution of  $\mathbf{Y}_i$  given  $b_i$ , and  $f_b(b_i | \mathbf{R}_i)$  is the conditional distribution of  $b_i$  given  $\mathbf{R}_i$ . The missing process  $\mathbf{R}_i$  affects the response  $\mathbf{Y}_i$  through the latent variable  $b_i$ .

We first present the model for the latent variable conditional on the missing process. It is assumed that  $b_i$  follows a normal distribution with unity variance and that values of the latent variable are affected by the missing status and other covariates. Under these assumptions, the latent variable  $b_i$  can be modelled through the equation

$$b_i = \mathbf{Z}_i^T \theta + \delta_i, \tag{4.2}$$

where  $\mathbf{Z}_i = (1, R_{i1}, R_{i2}, Z_{i1}, \dots, Z_{ip})$  is a  $(p+3) \times 1$  covariate vector,  $\theta$  is a  $(p+3) \times 1$  vector of coefficients and  $\delta_i \sim N(0, 1)$  is an error term. Then  $(b_i | \mathbf{R}_i) \sim N(\mathbf{Z}_i^T \theta, 1)$ .

Next, a model for the outcome conditional on the latent variables is specified. The complete response conditional on the latent variable is assumed to be normally distributed with mean and variance

$$E(Y_{ij} | b_i, \mathbf{W}_{ij}) = \mathbf{W}_{ij}^T \beta = \beta_0 + b_i \beta_1 + \mathbf{X}_{ij}^T \beta_2, \tag{4.3}$$

$$\text{var}(Y_{ij} | b_i, \mathbf{W}_{ij}) = \sigma_j^2, \tag{4.4}$$

for  $j = 1, \dots, K$ , where  $\beta = (\beta_0, \beta_1, \beta_2)$  is a vector of coefficients,  $\mathbf{W}_{ij} = (1, b_i, \mathbf{X}_{ij})$  is a vector of covariates for subject  $i$  at time  $j$  and  $X_{ij} = (X_{i1}, \dots, X_{iq})^T$ ,  $X_{i1}, \dots, X_{iq}$  are observed covariates. We assume no dependence between the variance, the covariance of  $Y_{i1}, \dots, Y_{iK}$  and latent variable  $b_i$ , but this assumption can be relaxed. An important assumption implied by model (4.3) is that, conditional on the latent variable, the missing data are MAR. That is, with the latent variable including in the model, the missingness no longer depends on

the missing data, after conditioning on observed data. This assumption is also used in Roy (2003) [8]’s paper, but it cannot be verified from the observed data. Given the assumptions stated above, maximum likelihood inference can be based on the distribution of the observed response vector  $Y_i$ , conditional on  $R_i$  and the covariates. In addition to the normal distribution, any manifest variables from an exponential family can be fit into the above latent variable mixed effect model.

#### 4.2.2 ESTIMATION

We use the EM algorithm to obtain parameter estimates. From equation (4.1), the log-likelihood that we want to maximize is

$$l(\beta, \sigma^2, \theta) = \sum_i \log \left( \int f_y(\mathbf{Y}_i | b_i; \beta, \sigma^2) f_b(b_i | \mathbf{R}_i; \theta) db_i \right), \quad (4.5)$$

where

$$f_y(\mathbf{Y}_i | b_i; \beta, \sigma^2) = \prod_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{1}{2\sigma_j^2} (Y_{ij} - \beta_0 - b_i\beta_1 - \mathbf{X}_{ij}^T\beta_2)^2\right],$$

and

$$f_b(b_i | \mathbf{R}_i; \theta) = \text{constant} \times \exp\left(-\frac{1}{2}(b_i - \mathbf{Z}_i^T\theta)^2\right).$$

*E-step.* We will obtain the expectation of functions of the data  $g(\mathbf{Y}_i, b_i)$  by the Hermite integration formula as Sammel, et al.(1997)[9] did,

$$E_b g(\mathbf{Y}_i, b_i) = \frac{\sum_{t=1}^T w_t \exp b_t^2 g(\mathbf{Y}_i, b_t) f_y(Y_i | b_t) f_b(b_t | R_i)}{\sum_{t=1}^T w_t \exp b_t^2 f_y(Y_i | b_t) f_b(b_t | R_i)} \quad (4.6)$$

The fixed abscissas  $b_t$  and corresponding weights  $w_t$  are obtained from the table for the Hermite intergrals (Abramowitz and Stegun, 1987[57]) (See Table 6) with  $T = 10$  being sufficient for accuracy.

*M-step.* We maximize the expected complete-data log-likelihood for the parameter  $(\beta, \sigma^2, \theta)$  at the M-step, which consists of the following steps: (1) Differentiating the log of  $f_y(\mathbf{Y}_i | b_i; \beta, \sigma^2)$  with respect to  $\beta$ , which gives

$$\frac{\partial \log f_y(\mathbf{Y} | b; \beta, \sigma^2)}{\partial \beta} = \sum_{i=i}^N \mathbf{W}_i (\mathbf{Y}_i - \mathbf{W}_i \beta). \quad (4.7)$$

Setting this quantity to be zero and taking an expectation with respect to the latent variable, gives

$$\hat{\beta} = \left( \sum_{i=1}^N E_b(\mathbf{W}_i \mathbf{W}_i^T) \right)^{-1} \sum_{i=1}^N E_b(\mathbf{W}_i) \mathbf{Y}_i. \quad (4.8)$$

(2) The partial derivative with respect to  $\sigma_j^2$  is

$$\frac{\partial \log f_y(\mathbf{Y}_{ij} | b_i; \beta, \sigma_j^2)}{\partial \sigma_j^2} = \sum_{i=1}^N \left[ -\frac{1}{2\sigma_j^2} + \frac{1}{2(\sigma_j^2)^2} (Y_{ij} - \mathbf{W}_{ij}^T \beta)^2 \right], \quad (4.9)$$

which implies that

$$\hat{\sigma}_j^2 = \frac{1}{N_j^*} \sum_i^N E_b(Y_{ij} - \mathbf{W}_{ij}^T \hat{\beta})^2, \quad (4.10)$$

where  $N_j^*$  is the number of observed continuous outcomes for the time point  $j$ .

(3) Differentiating  $\log f_b(b_i | \mathbf{R}_i; \theta)$  with respect to  $\theta$  gives

$$\frac{\partial \log f_b(b_i | \mathbf{R}_i; \theta)}{\partial \theta} = \mathbf{Z}_i (b_i - \mathbf{Z}_i^T \theta).$$

It now follows that

$$\hat{\theta} = \left( \sum_{i=1}^N \mathbf{Z}_i \mathbf{Z}_i^T \right)^{-1} \sum_{i=1}^N \mathbf{Z}_i E_b b_i, \quad (4.11)$$

where  $E_b b_i$  can be calculated by equation (4.6).

By beginning with reasonable initial estimates of the parameters, the E- and M-step are repeated by solving equations (4.8), (4.10) and (4.11) until differences in values of all the estimates in the consecutive iterations are sufficiently small.

The covariance matrix of the parameter estimates can be obtained by the bootstrap method or by directly calculating the inverse of the Fisher information matrix using the marginal log-likelihood of  $Y_{ij}$  at convergence. The marginal distribution of  $Y_{ij}$  is multivariate normal with mean and variance

$$E(Y_{ij}) = \mu_{ij} = \beta_0 + \mathbf{Z}_i^T \theta \beta_1 + \mathbf{X}_{ij}^T \beta_2, \quad (4.12)$$

$$\text{var}(Y_{ij}) = V_{ij} = \beta_1^2 + \sigma_j^2. \quad (4.13)$$

The first term in  $var(Y_{ij})$  comes from the variance of the latent variable  $b_i$ , and the second term comes from the conditional variance of  $Y_{ij}$  given  $b_i$ . So the marginal log-likelihood of  $Y_{ij}$  is

$$l_m = \sum_i \sum_j \left[ -\frac{1}{2} \log(\beta_1^2 + \sigma_j^2) - \frac{1}{2(\beta_1^2 + \sigma_j^2)} (Y_{ij} - \beta_0 - \mathbf{Z}_i^T \theta \beta_1 - \mathbf{X}_{ij}^T \beta_2)^2 \right].$$

Then, the negative values of twice derivatives with respect to  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$ ,  $\sigma_j^2$  and  $\theta$  are

$$\begin{aligned} -\frac{\partial^2 l_m}{\partial \beta_0^2} &= \sum_i \sum_j \frac{1}{V_{ij}}, \\ -\frac{\partial^2 l_m}{\partial \beta_1^2} &= \sum_i \sum_j \left[ \frac{(\mathbf{Z}_i^T \theta)^2 + 1}{V_{ij}} + \frac{4\beta_1 \mathbf{Z}_i^T \theta (Y_{ij} - \mu_{ij}) - (Y_{ij} - \mu_{ij})^2 - 2\beta_1^2}{V_{ij}^2} + \frac{4\beta_1^2 (Y_{ij} - \mu_{ij})^2}{V_{ij}^3} \right], \\ -\frac{\partial^2 l_m}{\partial \beta_2 \beta_2'} &= \sum_i \sum_j \frac{\mathbf{X}_{ij} \mathbf{X}_{ij}^T}{V_{ij}}, \\ -\frac{\partial^2 l_m}{\partial (\sigma_j^2)^2} &= \sum_i \left[ \frac{\beta_1^2 - \sigma_j^2}{V_{ij}^2} + \frac{(3\sigma_j^2 - \beta_1^2)(Y_{ij} - \mu_{ij})^2}{V_{ij}^3} \right], \\ -\frac{\partial^2 l_m}{\partial \theta \theta'} &= \sum_i \sum_j \frac{\mathbf{Z}_i \mathbf{Z}_i^T \beta_1^2}{V_{ij}}. \end{aligned}$$

The standard errors of the parameters can be obtained from the squared root of the inverse of the values (or the diagonal of the matrix) in the above equations.

### 4.3 APPLICATION TO THE KIDQUEST DATA

#### 4.3.1 DATA DESCRIPTIONS AND MODEL SPECIFICATIONS

To illustrate our method, we use a subset of data from the KidQuest study, which was conducted by Obesity/Nutrition Research Center, University of Pittsburgh. The subjects in this study are severely over-weight children who are 8-12 years old. One purpose of the KidQuest study is to show that children who participate in a family-based intervention program, have favorable changes in body mass index, body composition, food intake, activity level, and cardiovascular risk factors at 6-, 12- and 18-month assessments, when compared to those who receive usual care.



For our analysis we focus on the 133 children with complete baseline data. The entry criteria for the study are: to be 8.0 to 12.0 years of age; to have at least one parent or guardian willing to participate in the treatment program with the child; and to be  $\geq 150\%$  ideal body weight for height and age based on norms of the World Health Organization. Exclusion criteria were: mental retardation, pervasive developmental disorder or psychosis; genetic obesity syndrome; inability to engage in moderate exercise defined as 30 minutes of vigorous activity on most days of the week; or regular use of a medication that affects body weight or antidepressant medication, etc.

In this study, the continuous outcomes are a standardized measure of BMI, denoted  $Z_{BMI}$ . To calculate the standardized  $Z_{BMI}$ , we need the age and gender specific mean BMI ( $M_{BMI}$ ) and SD of that mean ( $SD_{BMI}$ ), which are obtained from Kuczmarski, et al. (2002) [58]. The formula for calculation of  $Z_{BMI}$  is

$$Z_{BMI} = \frac{BMI - M_{BMI}}{SD_{BMI}}.$$

Because the missingness of the data is informative, the missing indicators are related to the obesity outcomes. It is also assumed that the  $Z_{BMI}$ 's and the missing indicators are manifestations of an unobservable obesity severity score, the latent variable. Tables 7 and 8 give descriptive statistics for the  $Z_{BMI}$ 's and missingness. We see that compared with the usual care group, the children in the treatment group have a greater decrease in their  $Z_{BMI}$ 's at the 6-month assessment and that their  $Z_{BMI}$ 's increase thereafter. This is due to the fact that the treatment period is 6 months long and that during these 6 months their weights are controlled well, but after 6 months, their weights increase. The percentage of missingness increases from the 6-month to the 18-month assessment. In particular, the percentage of missingness is greatest for the 12-month assessment with the children in the usual care group. A possible reason is that the children in the usual care group have no change in  $Z_{BMI}$  at the 6-month assessment compared with the baseline, resulting in an increase in missingness at the next assessment.

Now we consider models for the latent variable and  $Z_{BMI}$ . Parameter estimates, standard errors and  $Z$ -values for these models are presented in Tables 10 – 11. In the model for the latent variable, we use a missing indicator ( $R_1$ ,  $R_2$ , see Table 9) and ‘treatment’ ( $1 =$

treatment, 0 = usual care) as the covariates. A negative estimate of  $R_1$  ( $\theta_1 = -0.0559$ ) and a positive estimate of  $R_2$  ( $\theta_2 = 0.1914$ ) imply that subjects who dropped out later and had fewer missing measurements before they dropped out had smaller values of the latent variable. In this case the latent variable can be treated as an unobservable obesity severity score, so that those subjects dropping out later and having fewer missing measurements before dropout had a smaller obesity severity scores. The relationship between the latent variable and the treatment is negative ( $\theta_3 = -0.1455$ ). Thus the children in the treatment group have a lower obesity score than the children in the usual care group. But the  $Z$ -value = -1.16 shows that the ‘treatment’ covariate is not significant. The reason for this is that the treatment was only applied for the first 6 months and in the remaining 12 months all of the children were under usual care.

Table 11 provides results for the modelling of  $Z_{BMI}$  for the proposed latent variable model, the pattern mixture model and GEE. In the latent variable model, the covariates are ‘intercept’, ‘latent variable’, ‘sex’, ‘time’ and ‘time<sup>2</sup>’. The ‘latent variable’ covariate is significant ( $\beta_1 = 2.0297$ ,  $Z$ -value = 4.39), so a larger  $Z_{BMI}$  is induced by a larger value of the latent variable. The ‘time’ and ‘time<sup>2</sup>’ covariates are also significant ( $\beta_3 = -0.0798$ ,  $Z$ -value = -3.62;  $\beta_4 = 0.0027$ ,  $Z$ -value = 2.25) with the interpretation that  $Z_{BMI}$  is lowest at 12 months. The estimate of the ‘sex’ covariate is positive and it is not significant ( $\beta_2 = 1.5434$ ,  $Z$ -value = 1.46). In both the pattern mixture model and GEE, the ‘treatment’ covariate is not significant which agrees with the latent variable model. But their estimates of the ‘sex’ covariate are positive and significant. It means that the boys have larger values of  $Z_{BMI}$  than the girls do. The latent variable model and GEE yield similar inference for significant ‘time’ and ‘time<sup>2</sup>’ effects, whereas the pattern mixture model does not show any significant ‘time’ and ‘time<sup>2</sup>’ effects. Table 12 gives the estimates of the conditional variances for  $Z_{BMI}$  at different time points in the latent variable model.

### 4.3.2 ASSESSING FIT OF THE MODEL

Here we use generalized Pearson residuals to detect outliers and assess the conditional independence assumption. The Pearson residual for outcome  $Y_{ij}$  can be calculated from

$$r_{ij} = \frac{Y_{ij} - E(Y_{ij})}{\sqrt{\text{var}(Y_{ij})}}, \quad (4.14)$$

where  $E(Y_{ij})$  and  $\text{var}(Y_{ij})$  are given in equations (4.12) and (4.13). The Pearson residuals versus the missing patterns (see Table 9) are plotted in Figure 2. It can be seen that almost all residuals fall within 2 units around 0, so our latent variable model fit the observed data quite well. Also there is no relationship between the residuals and the missing patterns, showing that the conditional independence assumption between the observed outcomes and missing patterns given the latent variable holds. Note that we cannot assess the fit of the model to the unobserved data and that we do not know whether the assumption of conditional independence between the unobserved outcomes and the missingness holds.

## 4.4 DISCUSSION

In this part of the dissertation, we proposed a latent variable model for longitudinal data with informative intermittent missingness. In our model, the value of the latent variable is determined by the missing patterns and it affects the observed outcomes. The proposed latent variable model is motivated by Roy (2003)[8]’s latent dropout class model. We consider the latent variable as continuous and assume that it is normally distributed with unity variance. In the real data, the latent variable commonly has its own meaning although it can not be observed directly. In the KidQuest study, the latent variable is considered as an unobservable obesity severity score.

The estimates of the parameters we are interested in are obtained by maximizing the log-likelihood and the EM algorithm is applied. One of the key operations in latent variable models is the summing or integrating over the latent variable. Because we assume that the latent variable is normally distributed, integration over the latent variable is possible. Here

we use the Gauss-Hermite quadrature as Sammel, et al. (1997)[9] did. Other methods for approximating the integral of the likelihood can also be considered, for example, Mauritsen (1990)[59] and Ten Have, et al. (1998)[18] approximate the normal integration by summing with respect to the binomial distribution. When the distribution of the latent variable is not simple (nonnormal), the integration becomes complicated.

To obtain the covariance matrix of the parameter estimates, the bootstrap technique or the inverse of the Fisher information matrix can be used. But when the outcomes are not normally distributed or the sample size is small, bootstrap technique or other asymptotic standard errors should be considered since they take account of the nonnormality of the observed outcomes in the maximum likelihood estimates.

Table 6: Hermite integration  $\int_{-\infty}^{\infty} g(x)dx = \sum_{i=1}^n w_i e^{x_i^2} g(x_i)$  for  $n = 10$ .

Abscissas( $\pm x_i$ )	Weight Factors ( $w_i$ )	$w_i e^{x_i^2}$
0.3429	$6.1086 \times 10^{-1}$	0.6871
1.0366	$2.4014 \times 10^{-1}$	0.7033
1.7567	$3.3874 \times 10^{-2}$	0.7414
2.5327	$1.3436 \times 10^{-3}$	0.8207
3.4362	$7.6404 \times 10^{-6}$	1.0255

Table 7: Descriptive statistics for  $Z_{BMI}$ .

Assessment	Treatment			Usual care			t-test	p-value
	n	mean	SD	n	mean	SD		
Baseline	68	5.35	2.02	65	5.39	1.73	-0.13	0.8967
6 months	55	4.27	1.65	48	5.36	1.76	-3.25	0.0016
12 months	49	4.76	2.27	34	4.93	1.64	-0.36	0.7211
18 months	42	4.41	1.68	41	4.87	1.58	-1.29	0.1999

Table 8: Descriptive statistics for missingness (frequency and percentage in the table are for the missingness).

Assessment	Treatment, n = 68		Usual care, n = 65		chi-sq. test (d.f. = 1)	p-value
	Frequency	%	Frequency	%		
6 months	13	19.12	17	26.15	0.94	0.3318
12 months	19	27.94	31	47.69	5.53	0.0187
18 months	26	38.24	24	36.92	0.02	0.8759

Table 9: Distribution of the missing patterns for KIDQUEST data

Pattern	$R_1$	$R_2$	Baseline	6 months	12 months	18 months	Frequency(%)
1	4	0	•	•	•	•	62 (46.62)
2	3	0	•	•	•	×	15 (11.28)
3	2	0	•	•	×	×	13 (9.77)
4	1	0	•	×	×	×	19 (14.29)
5	4	1	•	•	×	•	13 (9.77)
6	4	1	•	×	•	•	3 (2.26)
7	4	2	•	×	×	•	5 (3.76)
8	3	1	•	×	•	×	3 (2.26)
Total							133
•: Observed, ×: Missing							

Table 10: Estimates, estimated standard errors and Z-values for parameter of latent distribution

Parameter	Estimate	SE	Z-value
$\theta_0$ , intercept	1.2447	0.4904	2.54
$\theta_1$ , $R_1$	-0.0559	0.0648	-0.86
$\theta_2$ , $R_2$	0.1914	0.1738	1.10
$\theta_3$ , treatment	-0.1455	0.1253	-1.16

Table 11: Estimates, estimated standard errors and Z-values for modelling the outcomes,  $Z_{BMI}$ .

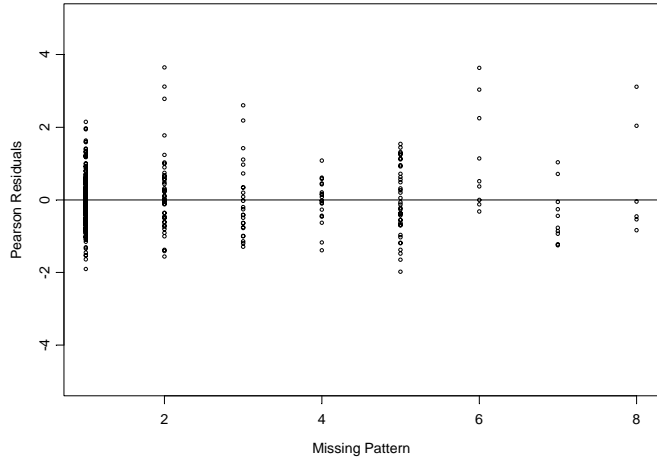
Variable	Proposed model			Pattern mixture model			GEE		
	Estimate	SE	Z-value	Estimate	SE	Z-value	Estimate	SE	Z-value
$\beta_0$ , intercept	4.1617	2.3463	1.77	4.6961	0.6243	7.52	5.1224	0.2602	19.69
$\beta_1$ , latent variable	2.0297	0.4621	4.39						
$\beta_1$ , treatment				-0.4738	0.6796	-0.70	-0.2513	0.2994	-0.84
$\beta_2$ , sex*	1.5434	1.0548	1.46	3.2169	1.1800	2.73	0.8707	0.3117	2.79
$\beta_3$ , time	-0.0798	0.0221	-3.62	-0.1416	0.2569	-0.55	-0.0799	0.0191	-4.18
$\beta_4$ , time <sup>2</sup>	0.0027	0.0012	2.25	0.0084	0.2288	0.04	0.0026	0.0010	2.60

\*sex: 1=male, 0=female.

Table 12: Estimates, estimated standard errors and Z-values for conditional variance of  $Z_{BMI}$  at different time points in the latent variable model

Parameter	Estimate	SE	Z-value
$\sigma_1^2$	0.7651	0.2392	3.20
$\sigma_2^2$	0.4292	0.1557	2.76
$\sigma_3^2$	0.6128	0.2480	2.47
$\sigma_4^2$	0.7495	0.2702	2.77





## 5.0 SUMMARY

We consider two methods based on latent variable models to analyze two longitudinal data sets with the outcome being subject to informative missingness. In the first approach, a latent class model is developed for binary outcomes with informative dropouts. This method can also be applied to intermittent missing data. The latent class model can be framed as a random effects model with the distribution of the random effects unspecified. The advantage of our proposed approach over a traditional random effects model is the simplicity of the implementation. Using this approach there are only two latent classes that are then added to the likelihood while the distribution of random effects must be considered for the random effects approach, which requires an assumption of normality and in most cases requires a more complicated likelihood involving integration of the random effects terms. To implement the proposed method, it is assumed that the population can be divided into two latent classes: a ‘regular’ class and a ‘special’ class. This assumption simplifies the model when compared with the other latent class models that need to determine how many latent classes are appropriate. However this special assumption may not hold in many data sets. The latent class model is also compared with weighted GEE (Robins, et al., 1995[43]), and a shared parameter model (Ten Have, et al., 1998[18]) in a simulation study and in an application to a real data set. In the simulation, three missing mechanisms corresponding to each of the three models are considered. It is not surprising that each model is the best one under its own missing mechanism, but the latent class model and the shared parameter model perform well under all the three missing mechanisms. The results show that the weighted GEE has a poor 95% coverage probability, especially under the latent class model missing mechanism. The proposed latent class model is the best one among these three models for

the simulation results in bias and 95% coverage probability. We also applied these three models to the women's smoking cessation data and the proposed latent class model and the shared parameter model have similar results. All of these results in the simulation and the application indicate that the proposed latent class model performs well for informative missing data and is a better choice when the data are appropriate for both the latent class model and the shared parameter model, since the calculation for the latent class model is simpler.

A pattern mixture model can also be very useful in a setting where the number of potential missing patterns is small, since the model specifically addresses the relationship between covariates and a given outcome within the framework of the missing patterns. However, this approach is very limited when there are many repeated measurements in a longitudinal study and/or when intermittent missingness is present. This is due to the fact that the number of missing patterns becomes too large, making implementation of the model difficult. To address this issue, we consider the latent variable model in the second part of this dissertation. This approach is an extension of the latent dropout class model by Roy (2003)[8], in which the outcomes and missing indicators are linked by a continuous latent variable. The model consists of continuous longitudinal outcomes, the missing indicators and the continuous latent variable. The value of the latent variable is affected by the missing pattern and it is also a covariate in modelling the outcomes. The parameter estimates are obtained by EM algorithm and the covariance matrix of the estimates can be approximated by using the bootstrap method or the inverse of the Fisher information matrix. This approach is then applied to the KidQuest data from Department of Psychiatry, University of Pittsburgh, and compared with the pattern mixture model and GEE. We use generalized Pearson residuals to detect the outliers and assess the fit of the proposed latent variable model.

## BIBLIOGRAPHY

- [1] P. J. Diggle, K. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, New York, 1994.
- [2] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley, New York, 1987.
- [3] R. J. A. Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83:1198–1202, 1988.
- [4] J. J. Heckman. The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimation method for such models. *Annals of Economic and Social Measurement*, 5:475–492, 1976.
- [5] R. J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88:125–134, 1993.
- [6] D. J. Bartholomew and M. Knott. *Latent Variable Models and Factor Analysis*. Arnold, London, 1999.
- [7] M. D. Perkins, K. A. and Marcus, M. D. Levine, D. D’Amico, A. Miller, M. Broge, and J. Ashcom. Cognitive-behavioral therapy to reduce weight concerns improves smoking cessation outcome in weight-concerned women. *Journal of Consulting and Clinical Psychology*, 69:604–613, 2001.
- [8] J. Roy. Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics*, 59:829–836, 2003.
- [9] M. D. Sammel, L. M. Ryan, and J. M. Legler. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society*, B 59:667–678, 1997.
- [10] K. Liang and S. Zeger. Longitudinal data analysis using generalized linear modes. *Biometrika*, 73:13–22, 1986.
- [11] F. Xie and M. C. Paik. Generalized estimating equation models for binary outcomes with missing covariates. *Biometrics*, 53:1458–1466, 1997.

- [12] J. S. Preisser, A. T. Galecki, K. K. Lohman, and L. E. Wagenknecht. Analysis of smoking trends with incomplete longitudinal binary responses. *Journal of the American Statistical Association*, 95:1021–1031, 2000.
- [13] S. R. Lipsitz, G. Molenberghs, G. M. Fitzmaurice, and J. Ibrahim. Gee with gaussian estimation of the correlations when data are incomplete. *Biometrics*, 56:528–536, 2000.
- [14] G. Y. Yi and R. J. Cook. Marginal methods for incomplete longitudinal data arising in clusters. *Journal of the American Statistical Association*, 97:1071–1080, 2002.
- [15] J. J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47:153–161, 1979.
- [16] M. C. Wu and R. J. Carroll. Estimation and comparison of changes in the presences of informative censoring by modeling the censoring process. *Biometrics*, 44:175–188, 1988.
- [17] D. Follmann and M. Wu. An approximate generalized linear model with random effects for informative missing data. *Biometrics*, 51:151–168, 1995.
- [18] T. R. Ten Have, A. R. Kunselman, E. P. Pulkstenis, and J. R. Landis. Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics*, 54:367–383, 1998.
- [19] P. S. Albert, D. A. Follmann, S. A. Wang, and E. B. Suh. A latent autoregressive model for longitudinal binary data subject to informative missingness. *Biometrics*, 58:631–642, 2002.
- [20] P. Diggle and W. G. Kenward. Informative drop-out in longitudinal data analysis (with discussion). *Applied statistics*, 43:49–93, 1994.
- [21] D. R. Cox. *The Analysis of Binary Data*. Chapman and Hall, London, 1970.
- [22] S. L. Zeger and B. Qaqish. Markov regression models for time series: A quasi-likelihood approach. *Biometrics*, 44:1019–1031, 1988.
- [23] E. Stasney. Some markov-chain model for nonresponse in estimating gross labor force flows. *Journal of Official Statistics*, 3:359–373, 1987.
- [24] M. R. Conaway. Non-ignorable non-response models for time-ordered categorical variables. *Applied Statistics*, 42:105–116, 1993.
- [25] B. F. Cole, M. T. Lee, G. A. Whitmore, and A. M. Zaslavsky. An empirical bayes model for markov-dependent binary sequences with randomly missing observations. *Journal of the American Statistical Association*, 90:1364–1372, 1995.
- [26] X. Liu, C. Waternaux, and E. Petkova. Influence of human immunodeficiency virus infection on neurological impairment: an analysis of longitudinal binary data with informative drop-out. *Applied statistics*, 48:103–115, 1999.

- [27] I. Deltour, S. Richardson, and J. Le Hesran. Stochastic algorithm for markov models estimation with intermittent missing data. *Biometrics*, 55:565–573, 1999.
- [28] P. S. Albert. A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics*, 56:602–608, 2000.
- [29] G. M. Fitzmaurice, N. M. Laird, and G. E. P. Zahner. Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, 91:99–108, 1996.
- [30] S. R. Lipsitz, J. G. Ibrahim, and G. M. Fitzmaurice. Likelihood methods for incomplete longitudinal binary response with incomplete categorical covariates. *Biometrics*, 55:214–223, 1999.
- [31] J. G. Ibrahim. Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85:765–769, 1990.
- [32] P. F. Lazarsfeld and N. W. Henry. *Latent Structure Analysis*. Houghton Mifflin, New York, 1968.
- [33] L. A. Goodman. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 62:215–231, 1974.
- [34] A. C. McCutcheon. *Latent Class Analysis*. Sage Publications, Beverly Hills, 1987.
- [35] B. Lindsay, C. C. Clogg, and J. Grego. Semiparametric estimation in the rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86:96–107, 1991.
- [36] A. Hadgu and Y. Qu. A biomedical application of latent class models with random effects. *Applied Statistics*, 47:603–616, 1998.
- [37] K. Roeder, K. G. Lynch, and D. S. Nagin. Modeling uncertainty in latent class membership: a case study in criminology. *Journal of the American Statistical Association*, 94:766–776, 1999.
- [38] C. C. Clogg. *Latent class models: Recent developments and prospects for the future*. In G. Arminger, C. C. Clogg and M. E. Sobel (Eds), *Handbook of Statistical Modelling for the Social and Behavioral Sciences*, pages 311–352. Plenum, New York, 1995.
- [39] E. S. Garret and S. L. Zeger. Latent class model diagnosis. *Biometrics*, 56:1055–1067, 2000.
- [40] B. A. Reboussin, M. E. Miller, K. K. Lohman, and T. R. Ten Have. Latent class models for longitudinal studies of the elderly with data missing at random. *Applied Statistics*, 51:69–90, 2002.

- [41] B. A. Reboussin, D. M. Reboussin, K. Y. Liang, and J. C. Anthony. A latent transition approach to modeling progression of health-risk behavior. *Multivariate Behavioral Research*, 33:457–478, 1998.
- [42] B. A. Reboussin, K. Y. Liang, and D. M. Reboussin. Estimating equations for a latent transition model with multiple discrete indicators. *Biometrics*, 55:839–845, 1999.
- [43] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90:106–121, 1995.
- [44] H. Lin, B. W. Turnbull, C. E. McCulloch, and E. H. Slate. Latent class models for joint analysis of longitudinal biomarker and event process data: Application to longitudinal prostate-specific antigen readings and prostate cancer. *Journal of the American Statistical Association*, 97:53–65, 2002.
- [45] B. H. Patterson, C. M. Dayton, and B. I. Graubard. Latent class analysis of complex sample survey data: Application to dietary data. *Journal of the American Statistical Association*, 97:721–741, 2002.
- [46] S. le Cessie and J. C. van Houwelingen. Logistic regression for correlated binary data. *Applied Statistics*, 43:95–108, 1994.
- [47] R. L. Prentice. Correlated binary regression with covariance specific to each binary observation. *Biometrics*, 44:1033–1048, 1988.
- [48] J. R. Ashford and R. R. Sowden. Multivariate probit analysis. *Biometrics*, 26:535–546, 1970.
- [49] L. J. Emrich and M. R. Piedmonte. A method for generating high-dimensional multivariate binary variates. 45:302–304, 1991.
- [50] D. J. Ossip-Klein, G. Bigelow, S. R. Parker, S. Curry, S. Hall, and S. Kirkland. Classification and assessment of smoking behavior. *Health Psychology*, 5 (Suppl.):3–11, 1986.
- [51] D. B. Rubin. Inference and missing data. *Biometrika*, 63:581–590, 1976.
- [52] W. Guo, S. J. Ratcliffe, and T. T. Ten Have. A random pattern-mixture model for longitudinal data with dropouts. *Journal of the American Statistical Association*, 99:929–937, 2004.
- [53] C. C. Clogg. Some latent structure models for the analysis of likert-type data. *Social Science Research*, 8, 1979.
- [54] D. Rindskopf and W. Rindskopf. The value of latent class analysis in medical diagnosis. *Statistics in Medicine*, 5:21–27, 1986.

- [55] I. Moustaki and M. Knott. Weighing for item non-response in attitude scales by using latent variable models with covariates. *Journal of Royal Statistical Association*, 163:445–459, 2000.
- [56] H. Lin, C. E. McCulloch, and R. A. Rosenheck. Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics*, 60:295–305, 2004.
- [57] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover Publications, New York, 1987.
- [58] R. J. Kuczmarski, C. L. Ogden, S. S. Guo, and et al. 2000 cdc growth charts for the united states: methods and development. national center for health statistics. *Vital and Health Statistics Series Reports*, 11(246):147–148, 2002.
- [59] R. Mauritsen. *Egret Reference Manual*. Statistics and Epidemiology Research Corporation, Seattle, 1990.