The Pennsylvania State University

The Graduate School

Department of Statistics

A LATENT-CLASS SELECTION MODEL FOR NONIGNORABLY MISSING DATA

A Thesis in

Statistics

by

Hyekyung Jung

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2007

The thesis of Hyekyung Jung has been reviewed and approved* by the following:

Joseph L. Schafer
Associate Professor of Statistics
Thesis Advisor
Chair of Committee

John W. Graham
Professor of Biobehavioral Health

Aleksandra B. Slavković
Assistant Professor of Statistics

Bruce G. Lindsay
Willaman Professor of Statistics
Head of the Department of Statistics

*Signatures are on file in the Graduate School

Abstract

A Latent-Class Selection Model

for Nonignorably Missing Data

Most missing-data procedures assume that the missing values are ignorably missing or missing at random (MAR), which means that the probabilities of response do not depend on unseen quantities. Although this assumption is convenient, it is sometimes questionable. For example, questionnaire items pertaining to sensitive information (e.g., substance use, delinquency, etc) may show high rates of missingness. Participants who fail to respond may do so for a variety of reasons, some of which could be strongly related to the underlying true values.

Data are said to be nonignorably missing if the probabilities of missingness depend on unobserved quantities. Traditional selection models for nonignorable nonresponse are outcome-based, tying these probabilities to partially observed values directly (e.g., by a logistics regression). These methods are inherently unstable, because the relationship between a partially observed variable and its missingness indicator is understandably difficult to estimate. Moreover, with multivariate or longitudinal responses, the number of distinct missingness patterns becomes quite large, making traditional selection modeling even more unattractive. Information in the missing-data indicators is sometimes well summarized by a simple latent-class structure, suggesting that a large number of missing-data patterns may be reduced to just a few prototypes.

In this thesis, we describe the new method for imputing missing values under a latent-class selection model (LCSM). In the LCSM, the response behavior is assumed to be related to the items in question, and to additional covariates, only through a latent membership measured by the missingness indicators. We describe the LCSM and apply it to data from a school-based study of alcohol risk and exposure among adolescents in Pennsylvania, which has sensitive items with high rates of missingness. We examine alcohol risk index for students from 8 to 13 years old and compare our model's performance to that of MAR-based alternative.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my deepest appreciation to my wonderful mentor and advisor Dr. Joseph Schafer for his guidance, training and support throughout graduate school and the research process. This thesis would not have been possible without his help. It is my honor to have him as my advisor, to learn and get inspiration from him. Also, I am grateful to the valuable suggestions and comments from Dr. Bruce Lindsay, Dr. Thomas Hettmansperger, Dr. John Graham, and Dr. Aleksandra Slavković on this research and their presence in my committee.

Many thanks go to the people in the Methodology Center of the Pennsylvania State University, for the support and various perspectives and points of view. I also want to thank my dearest parents and brothers, for their endless love and support. Lastly, and most importantly, I wish to thank my husband, Byungtae Seo, who have given me a great deal of help as a colleague throughout the PhD program. My daughter, who will be born soon, has added great pleasure to the process of producing this work. To them I dedicate this thesis.

Chapter 1

Introduction

1.1 Multivariate Incomplete Data

Statistical analysis and modeling of multivariate data with missing values poses many challenges. Early work on this problem focused on efficient estimation of means and covariance matrices. Covariances are important because they provide a basis for multiple regression, principal component analysis, structural equation modeling, and so on. Literature on multivariate missing-data methods is reviewed by Little and Rubin (2002), and Schafer (1997). The Expectation Maximization (EM) algorithm (Beale and Little, 1975; Dempster et al., 1977) produces maximum likelihood (ML) estimates of means and covariances under an assumption of normality (Little and Rubin, 2002, Chapter 8). A useful alternative to ML methods is multiple imputation (MI) (Rubin, 1987; Schafer, 1997). In MI, each of the missing values is replaced by a set of plausible simulated values, which represent uncertainty about the missing data. Whereas the EM algorithm produces estimates of model-specific parameters, a multiply imputed dataset can be analyzed in a variety of ways. The resulting completed datasets can be used by multiple researchers. EM and MI have also been implemented under non-normal models, including log-linear models for multivariate categorical data, the general location model for mixed datasets containing both continuous and categorical data, and a multivariate linear mixed–effects model for multivariate panel data or clustered data (Schafer, 1997, Chapter 2).

The approaches mentioned above are almost invariably implemented under the assumption that the missing values in the dataset are missing at random (MAR) (Rubin, 1976). MAR essentially means that the probabilities of missingness may depend on observed data, but are conditionally independent of all missing values given the observed ones. However, this assumption is often questioned. Reasons for missing values are often thought to be related to the values themselves. For instance, individuals may refuse to answer sensitive items (e.g., pertaining to income or drug use) on a questionnaire for reasons related to the underlying true values for those items.

In multivariate settings with arbitrary patterns of missingness, the MAR assumption is mathematically convenient, but it is intuitively unappealing and often implausible (Robins and Gill, 1997). In a multivariate setting, MAR means that a subject's probabilities of responding to items may depend only on his or her own set of observed items, a set that changes from one subject to the next, which seems odd or unnatural.

If we suspect that missingness may depend on missing values, then a proper analysis requires us to jointly model the population of the complete data and the missingness indicators. Many articles have been published on this problem, particularly in the context of longitudinal studies with dropout (Diggle and Kenward, 1994; Ibrahim et al., 2001; Little, 1995; Troxel et al., 1998). Dropout produces a missing-data pattern that is monotone, in the following sense: Suppose Y_{ij} is the measurement for subject i at occasion j. Missingness is montone if, whenever an element Y_{ij} is missing, Y_{ik} is also missing for all $k > j$ (Rubin, 1974). Missingness patterns in longitudinal studies are often monotone or nearly so, because once a subject drops out of the study in a given wave, the subject usually does not return in subsequent waves.

Relatively little has been published on non-MAR non-monotone missingness for

general multivariate problems. Monotone missingness can be described with a simple sequence of models predicting the probability that a subject drops out at occasion $j + 1$ given that the subject is still present at occasion j. A multivariate dataset with p variables, however, may have as many as 2^p possible patterns. In that case, modeling the relationships among the missingness indicators and their relationships to the incomplete data is challenging. Some work has been done on this problem with categorical responses (Park, 1998; Fitzmaurice et al., 1996a). For example, Rubin et al. (1995) and Molenberghs et al. (2001) presented analyses of survey data related to the Slovenian plebiscite, allowing the probabilities of missingness to depend on the items in various ways. This example, however, had only a few variables. Relatively little has been done on the problem of nonignorable nonresponse where the underlying population is assumed to be multivariate normal. Little (1993, 1994) explored pattern-mixture models for multivariate missing data, and Scheid (under review) proposes a selection model for bivariate normal distributed data with nonignorable nonresponse. Once again, however, these models are practical for only a small number of variables or patterns.

If missingness is systematically related to outcomes of interest, and if these non-MAR aspects of the data are not taken into account in the analysis, the resulting estimates of population parameters may be biased (Pirie et al., 1988; Vach and Blettner, 1995). Moreover, the results of the study may be difficult to generalize, because the respondents may not represent the target population of interest, again due to differential rates of missingness across different types of subjects.

In practice, investigators can never be sure whether departures from the MAR assumption in their data are severe enough to make a difference. Even if the primary analysis proceeds under an assumption of MAR, it is worthwhile to investigate how the results may change under different assumptions. A standard ignorable analysis can be

strengthened by sensitivity analyses that include nonignorable alternatives. Results become more convincing if estimates from a variety of alternative models agree. If they do not agree, the differences impart a better sense of the true levels of uncertainty.

Nonignorable models that have been proposed thus far have tended to be problem-specific and do not generalize well. The primary goal of this research is to develop a general method for nonignorable modeling of incomplete multivariate data based on the idea of latent class modeling (Goodman, 1974; McCutcheon, 1987). We will summarize the distribution of the missingness indicators through a latent-class model, and then relate subjects' latent-class memberships to the variables containing the missing values. A detailed description of this new model will be given in Chapter 3.

1.2 Notations and Definitions

Some notational conventions will be used throughout this thesis. We will use Y_{ij} to denote the response of the ith subject to the jth variable. For notational ease, we will use unbolded-Y to denote a vector as well as a variable. A complete dataset will be denoted by a matrix Y_{com} with n rows and p columns, where n represents the number of subjects and p represents the number of variables. We will also denote the observed portions of Y_{com} as Y_{obs}, and the unobserved items as Y_{mis}, so the complete data Y_{com} can be written as $Y_{com} = (Y_{obs}, Y_{mis})$. This partitioning of Y_{com} can be encoded in a set of random variables R, a matrix with the same dimensions as Y_{com}, whose elements take the value of 1 if the corresponding element of Y_{com} is observed and 0 if the element of Y_{com} is missing. R will be called the missingness indicators. $P(R \mid Y_{com}; \xi)$ is the missingness mechanism, or the distribution of missingness, which specifies how the probabilities of missingness are related to the complete data, and ξ represents parameters of the missingness mechanism. These relationships should be

interpreted in a correlational, rather than a causal, sense. A model for the distribution of missingness does not need to include the "true causes" of R; it only needs to accurately describe the relationship between R and Y_{com} in the population.

A useful classification of missingness mechanisms is presented by Little and Rubin (2002, Chapter 2):

1. Missing data are said to be missing completely at random (MCAR) if the distribution of missingness does not depend on Y_{obs} or Y_{mis}, $P(R \mid Y_{com}; \xi) = P(R; \xi)$.

2. Missing data are said to be missing at random (MAR) if the distribution of missingness does not depend on Y_{mis}, $P(R \mid Y_{com}; \xi) = P(R \mid Y_{obs}; \xi)$.

3. Missing data are missing not at random (MNAR) if there is any violation of MAR, $P(R \mid Y_{com}; \xi) \neq P(R \mid Y_{obs}; \xi)$.

The third situation is often termed "nonignorable" and is the main focus of this research. MNAR means that, even after accounting for all the available observed information, the reasons for observations being missing still depend on the unseen observations themselves. To obtain valid inferences under MNAR, a joint model for the complete data Y_{com} and the missingness indicators R is required. Unfortunately, one can never tell from the data at hand whether the missing values are MAR or NMAR. Assessing the plausibility of MAR versus MNAR requires external knowledge of the subject matter. Rather than trying to judge whether the missing values are MAR, we will develop tools to explore how inferences may change if the assumptions of MAR is violated.

1.3 Motivating example

The methodology in this thesis was partly motivated by data from a school-based study of alcohol risk and exposure among adolescents in the Pennsylvania area.

The data were collected as part of pilot study to investigate the characteristics of a new instrument called Assessment of Liability and Exposure to Substance Use and Antisocial Behavior (ALEXSA) (Ridenour et al., under review). This instrument was designed to measure predictors of addiction and other problem behaviors (APBs) in students aged 8 to 13. One section of ALEXSA was devoted to measures of alcohol exposure and risk, to estimate the levels of social and environmental exposure to alcohol as a predictor of APBs. Students who correctly recognized alcohol from a given set of pictures were presented a series of alcohol-related questions. These questions pertained to how they first learned about alcohol, parents'/friends'/own attitudes regarding alcohol, normative beliefs about alcohol use by persons at the respondent's age, ever being offered/ever tried the drug, availability of alcohol, expected high school use, and frequency of current use. These items were designed to be aggregated into a single overall measure of alcohol risk and exposure.

In the ALEXSA pilot study, these alcohol-related items contain a mixture of planned and unplanned missing values. Planned missingness arises from the initial question, which we called the stem question, which determines whether the subject knows what alcohol is. If so, then he or she proceeds to the remaining items. However, there are some missing values on the stem question, because data collectors could not always determine whether the subject recognized alcohol. Additionally, coding problems sometimes led data collectors to skip the subsequent items when the subject really did know what alcohol is. Finally, even when subjects did know and were directed to the remaining items, some of their responses to those items were "don't know" or refused. The researchers suspected that these missing values were often related to the underlying answer. A subject may have responded with a "don't know" or refusal to mask a true answer that was socially undesirable. Given the potential for nonignorable missingness, the researchers thought it would be important to consider

alternative assumptions about missingness, to investigate the possibilities for bias and for possibly misleading conclusions from an analysis based on MAR alone. More detailed information about ALEXSA dataset will be given in Chapter 5.

The rest of this thesis is organized as follows. Chapter 2 provides a review and a discussion of published work related to the problem of multivariate incomplete data with nonignorable missingness. Two types of nonignorable models—selection and pattern-mixture models—are described, along with their advantages and limitations.

Chapter 3 presents some general notation and definitions for latent-class (LC) models. We then present our new model, which we call a latent-class selection model (LCSM), for nonignorably missing multivariate data. In Chapter 4, we describe computational strategies for Bayesian inference and multiple imputation under the LCSM. We also discuss issues of model checking and selection of prior distributions for the LCSM parameters.

In Chapter 5, we first apply the LCSM to a small example dataset, the Foreign Language Attitude Scale (FLAS) data described by Schafer (1997), to show that missingness indicators can often be well described by a simple latent class structure. We then apply the LCSM to our motivating example from ALEXSA, using the computational algorithms described in Chapter 4, and compare the results from the LCSM to those from an ignorable model.

Finally, we provide an in-depth discussion of features, possible extensions of our LCSM, conclusions and future work in Chapter 6.

Chapter 2

Overview of the Literature on Nonignorable models for Incomplete Multivariate Data

If there is reason to suspect that missingness may depend on missing observations themselves, so the MAR assumption is questionable, alternative procedures may be developed by proposing models for the missingness mechanism that relate probabilities of missingness to the missing values. The most common approach is to construct a fully parametric model for the joint distribution of the complete data and missingness indicators, and then estimate the unknown parameters using either maximum likelihood or Bayesian methods (Little and Rubin, 2002, Chapter 8).

Let R be the missingness indicators which separate the complete data Y_{com} into (Y_{obs}, Y_{mis}). Let $P(Y_{com}, R \mid \theta, \xi)$ be the joint distribution for the complete data and the missingness indicators, where θ and ξ are sets of unknown parameters that characterize the joint distribution. Because Y_{mis} is not seen, the evidence about θ and ξ is summarized by the likelihood function given by the integral of this distribution over the unseen missing values,

$$\begin{aligned} L(\theta, \xi \mid Y_{obs}, R) &\propto P(Y_{obs}, R \mid \theta, \xi) \\ &= \int P(Y_{obs}, Y_{mis}, R \mid \theta, \xi)\, dY_{mis}. \end{aligned} \tag{2.1}$$

If any unmodeled covariates are present, conditioning on covariates will be implicit in the notation. The practical implication of missing not at random (MNAR) is that

the likelihood requires an explicit model for R. When Y_{mis} is MAR, a model for R is not needed. In that case, inferences about the population distribution of Y_{com} may be based on a simpler likelihood function that integrates Y_{mis} out of the marginal distribution of Y_{com}.

From this likelihood standpoint, there are two major approaches to construct nonignorable models based on different factorizations of the joint distribution of Y_{com} and R: selection models (Diggle and Kenward, 1994; Heckman, 1976) and pattern–mixture models (Little, 1993).

2.1 Selection Models

Selection models, which were introduced by the econometrician Heckman (1976), factor the joint distribution of the complete data and the missingness indicators into a marginal density for the complete data and a conditional density for the missingness indicators given the complete data,

$$P(Y_{com}, R \mid \theta, \xi) = P(Y_{com} \mid \theta)\ P(R \mid Y_{com}, \xi). \tag{2.2}$$

This factorization is conceptually attractive because the parameters of primary interest, θ, which describe population of complete data, appear in the first term, and the second term is the missing-data mechanism that appears in the definition of MAR. Heckman applied these models to a univariate response, and Diggle and Kenward (1994) extended them to multivariate responses. Fitzmaurice et al. (1996b) developed selection models for incomplete binary responses under a multivariate logistic distribution for the population, and Molenberghs et al. (1997) applied similar methods to longitudinal ordinal data. The most widely used selection models today combine a generalized linear regression for a response given covariates with a logit or probit regression for missingness given the covariates and the response. Selection models

for cross-sectional data have been implemented in the software packages LIMDEP (Greene, 1991) and aML (Lillard and Panis, 2000). The OSWALD package (Smith et al., 1996), based on the extension of the work in Diggle and Kenward (1994), can fit selection models for longitudinal data.

Selection models are intuitively appealing because they allow researcher to formalize in the second term $P(R \mid Y_{com}, \xi)$ their notions of how the probabilities of missingness depend directly on the data values. However, selection models should be approached with caution. Untestable restrictions must be placed on the missingness mechanism to make these models identifiable (Glynn et al., 1986). Results from selection models can be highly sensitive to different assumptions about the shape of the complete data population (Little and Rubin, 2002, Chapter 15). With continuous responses, the common assumption of a normally distributed population cannot be verified when some responses are missing (Hogan et al., 2004). Slight perturbations to the population model—e.g., assuming a Student's t-distribution for the population rather than a normal—may cause drastic changes in parameter estimates (Kenward, 1998). These models are also sensitive to the functional form of the relationship between the missingness indicators and the complete data. Tying probabilities of missingness to the partially observed values directly (e.g., by logit or probit regression) can make estimates unstable (Diggle and Kenward, 1994).

Parameter estimates for selection models are usually obtained by maximizing the likelihood function (2.1). The loglikelihoods for these problems are often oddly shaped. The surface may be nearly flat with respect to some aspects of ξ, leading to numerical instability (Hogan and Laird, 1997).

2.2 Pattern-Mixture Models

A popular alternative to selection modeling is to factor the joint distribution of the complete data and the missingness indicators into a marginal distribution for the missingness indicators and a conditional distribution for the complete data given the pattern of missingness,

$$P(Y_{com}, R \mid \theta, \xi) = P(R \mid \xi)\ P(Y_{com} \mid R, \theta). \tag{2.3}$$

These are called pattern-mixture models (Little, 1993, 1994, 1995; Little and Wang, 1996). The parameters in this approach have a different meaning from those in selection models. Parameters describing the marginal distribution of Y_{com} do not appear in the factorization (2.3), but must be obtained by manipulation of ξ and θ. Some researchers find these less intuitively appealing than selection models, because they are more accustomed to thinking about how R is influenced by Y_{com}, rather than how Y_{com} depends on R. Pattern-mixture models have computational advantages, however, because likelihood functions of the form

$$\begin{aligned} L(\theta, \xi \mid Y_{obs}, R) &\propto P(Y_{obs}, R \mid \theta, \xi) \\ &= \int P(Y_{obs}, Y_{mis} \mid R, \theta) P(R \mid \xi)\ dY_{mis} \end{aligned} \tag{2.4}$$

tend to be easier to work with than those from selection models. Moreover, parameters that cannot be estimated from the joint distribution of R and Y_{obs} are more readily identified in the pattern-mixture framework than in the selection framework (Little, 1993).

Pattern-mixture models describe the population of the complete data as a mixture of distributions, weighted by the marginal proportions of subjects in the various missingness patterns. Marginalization over the patterns is usually required to obtain parameter estimates of primary interest, which pertain to the population of Y_{com}. In

this approach, we stratify the incomplete data by missingness patterns, fit distinct models within each stratum, and aggregate the results over patterns.

The underlying assumption of pattern-mixture modeling is that every subject with the same missingness pattern shares a common distribution. When the number of unique missingness patterns is large, the observations within many strata become sparse, and parameters estimates from those strata may be unstable. Moreover, the observed data within a pattern give no information about the aspects of θ that pertain to the conditional distribution of the missing values given the observed values within that pattern. In order to estimate θ, identifying restrictions must be placed on the parameter space (Wu and Bailey, 1984; Little, 1995; Daniels and Hogan, 2000).

Multivariate responses can lead to a large number of patterns, and fitting separate models to each pattern becomes a daunting task. In practice, the patterns are often grouped together. In a clinical trial, Heddeker and Gibbons (1997) classified subjects into just two groups: those who completed the six-week trial, and those who dropped out at any time prior to the final measurement. Coarsening the information in R in this way can simplify the construction of a pattern-mixture model. Roy (2003) proposed a pattern-mixture model that grouped subjects into a small number of latent classes. His model is related to the new model proposed in this thesis; the differences will be described in Chapter 3.

2.3 Related Work

Little and Rubin (2002, Chapter 15) provide a general discussion and examples of nonignorable missing-data models. Little (1995) gives a detailed review of pattern–mixture and selection models for longitudinal studies, and he characterizes general classes of models for nonignorable dropout. MNAR dropout is also discussed by Hogan and Laird (1997), Kenward and Molenberghs (1999), and Verbeke and Molenberghs

(2000). Ibrahim et al. (2005) examine four common approaches for inference (ML, MI, fully Bayesian, and weighted estimating equations) in generalized linear modeling with selection models for the missing-value process. Articles on nonignorable missing-data models for survey data have been written by Stasny (1987, 1988, 1990), Conaway (1992, 1993), Chambers and Welsh (1993), Forster and Smith (1998), and Heitjan and Landis (1994).

Model checking and criticism can be challenging with incomplete data, and especially so when the model assumes the missing values are MNAR. Alternative functional forms for the missingness mechanism can be compared by the likelihood ratio or the Akaike Information Criterion (AIC). In practice, however, the observed data usually provide little or no information to distinguish among alternative nonignorable models (Demirtas and Schafer, 2003). Many authors have stressed the central role of sensitivity analysis, in which results from a variety of models are compared. Sensitivity analyses for MNAR missing-data models are discussed by Verbeke et al. (2001), Fairclough et al. (1998), Baker et al. (2003), and Michiels et al. (1999).

The fundamental challenge arising in nonignorable modeling is parameter identification (Baker and Laird, 1988). Certain aspects of the joint distribution of Y_{com} and R will never be estimated from the quantities that are seen, which are Y_{obs} and R. The challenge is to create a model that applies information that is strong enough to identify the parameters, yet weak enough to allow the data to speak for themselves and accurately reflect uncertainty. The different factorizations of the likelihood used in selection and pattern-mixture models naturally lead to different types of identifying restrictions. In selection models, the restrictions are placed on $P(R \mid Y_{com}, \xi)$, whereas in pattern-mixture models they are applied to $P(Y_{com} \mid R, \theta)$. Other kinds of restrictions can be imposed by introducing latent variables that attempt to capture the relationships between Y_{com} and R in a parsimonious way. Wu and Carroll (1988)

and Wu and Bailey (1988, 1984) allowed responses to depend on missingness indicators through individual random effects estimated from a general linear mixed model, and then averaged over the distribution of the random effects. Their model can be written as

$$P(Y_{com}, R \mid \theta, \xi) = \int P(R \mid b, \xi) P(Y_{com} \mid b, \theta)\, dF(b), \tag{2.5}$$

where $F(b)$ is a distribution for the subject-specific random effects. In this model, Y_{com} and R are linked through b, and inferences are based on the likelihood obtained by integrating (2.5) over b and Y_{mis}. This is an example of what is often called a shared parameter model. The shared parameter, b, is a latent trait that drives both the measurement and missingness processes. Follmann and Wu (1995) extended this idea to permit generalized linear model for discrete responses with no parametric assumptions on the distribution of random effects. In a longitudinal setting, Thijs et al. (2002) allowed different missing-data patterns to share certain parameters so that the patterns with less data could borrow information from patterns with more data.

The natural parameters of selection models, pattern-mixture models, and shared parameter models have very different meanings, and transforming one kind of model into another is not straightforward. Directly comparing the results from models with different parameterizations can be difficult. For sensitivity analyses that span multiple types of models, Demirtas and Schafer (2003) propose the use of multiple imputation (MI) (Rubin, 1987) in which values of Y_{mis} are repeatedly simulated from a posterior predictive distribution given Y_{obs} and R. After imputation, all information about the missingness mechanism is carried in the imputed values, and imputed datasets from different models can be analyzed in exactly the same way.

Chapter 3

A Latent-Class Selection Model

3.1 Traditional Latent-Class Models and Latent-Class Regression

Given a set of categorical measurements on a sample of units, a researcher may wish to know if the structure of the data can be explained by classifying units into a small number of groups of clusters. Latent-class (LC) modeling is one method for identifying groups of similar units. LC models explain the relationships among the observed categorical variables or items by an unseen (i.e., latent) classification whose membership is inferred from the data. These models have been used by psychiatrists to classify persons into diagnostic categories given the presence/absence of multiple symptoms. LC models have been applied to survey data by sociologists and marketing researchers to identify subgroups of the population holding particular attitudes or preferences. Researchers in psychiatry have used them as an alternative to traditional item-response theory (IRT) models, which measure subjects' abilities on a continuous scale. LC models are more appropriate than IRT when the researchers are trying to identify subjects who have understood or mastered a task or concept. General overviews of LC modeling are provided by Goodman (1974), Haberman (1979), Clogg (1995), and McCutcheon (1987).

Let $Y_i = (Y_{i1}, \ldots, Y_{ip})$ denote a vector of p polytomous items for the ith subject, where Y_{ij} takes possible values $1, 2, \ldots, M_j$. These variables may be nominal or ordinal, but we will not take ordering into account in this description of the LC model. We

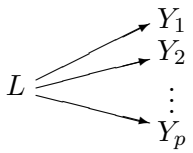

Figure 3.1: Relationship among latent variable and items in the LC model

will suppose that a latent variable exists which, if it were seen, would explain the relationships among the items. Let L_i denote the latent variable, which will take values $1, 2, \ldots, C$. LC models assume that the observed items in Y_i are mutually independent within each class of L_i. This assumption of conditional independence is called "local independence" (Lazarsfeld and Henry, 1968). It supposes that, once the effect of latent class membership on the data is taken into account, all that remains is random noise. Similar assumptions of conditional independence are used in factor analysis, IRT modeling and other tools used by social and behavioral scientists to account for measurement error.

In practice, the assumption of local independence is sometimes violated. For example, a questionnaire may have two items that are so similar that responses to them may be strongly associated among individuals in a given latent class. In these situations, the traditional LC model can be extended to a more general class of log-linear models, which will allow more complex associations among the augmented set of variables $(Y_{i1}, \ldots, Y_{ip}, L_i)$ (Haberman, 1979; Hagenaars, 1990, 1993; Uebersax and Grove, 1990). These more complex models are rarely used in practice, however, and in this thesis we will restrict our attention to the model of local independence. The relationships among the items and latent-class variable assumed by the standard LC model are shown in Figure 3.1.

Under local independence, the probability of a pattern of item responses in

given class $L_i = l$ can be written as

$$P(Y_{i1} = y_1, \ldots, Y_{ip} = y_p \mid L_i = l) = \prod_{j=1}^{p} \prod_{m=1}^{M_j} P(Y_{ij} = m \mid L_i = l)^{I(Y_{ij}=m)}, \qquad (3.1)$$

where $I(Y_{ij} = m)$ denotes an indicator function which takes the value 1 if Y_{ij} is equal to m and 0 otherwise. The probability of a pattern of item responses regardless of class membership is then

$$\begin{aligned} P(Y_{i1} = y_1, \ldots, Y_{ip} = y_p) &= \sum_{l=1}^{C} P(L_i = l) \prod_{j=1}^{p} \prod_{m=1}^{M_j} P(Y_{ij} = m \mid L_i = l)^{I(Y_{ij}=m)} \\ &= \sum_{l=1}^{C} \pi_l \prod_{j=1}^{p} \prod_{m=1}^{M_j} \rho_{jm|l}^{I(Y_{ij}=m)}, \end{aligned} \qquad (3.2)$$

where π_l is the prevalence of class l in the population, and $\rho_{jm|l}$ is the probability that a member of class l responds to item Y_{ij} with a value of m. The LC model is a finite mixture of discrete multivariate distributions. Finite-mixture models pose many interesting challenges for parameter estimation and inference because of their unusual geometry (Lindsay, 1995; Titterington et al., 1985). For example, the likelihood functions for finite mixtures are invariant to permutations of the class labels.

The class prevalences in the LC model will be written as $\pi = (\pi_1, \ldots, \pi_C)$, and the item-response probabilities (which are also called "measurement parameters") will be denoted by $\rho = (\rho_{11|1}, \ldots, \rho_{1M_1|1}, \rho_{21|1}, \ldots, \rho_{pM_p|C})$. Maximum-likelihood (ML) estimates of the parameters are sometimes computed by Fisher scoring or Newton-Raphson. The most popular method, however, is the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). EM is a natural choice for LC models and other finite-mixtures due to its simplicity and stability. The application of EM to LC models is discussed by Goodman (1974), Hagenaars (1990) and Bartholomew and Knott (1999).

After an LC model is built to identify groups in population, it is natural to extend the model to relate class membership to covariates. Extensions of the tradi-

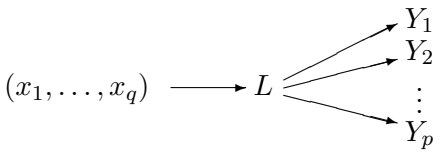

Figure 3.2: Relationship among latent variable, covariates and items in the LC regression model

tional LC model to allow class membership probabilities π to depend on covariates were developed by Clogg and Goodman (1984), Dayton and Macready (1988), and Bandeen-Roche et al. (1997).

Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{iq})^T$ denote a vector of covariates for ith subject that may influence his or her probabilities of belonging to the latent classes $L_i = 1, \ldots, C$. In most cases, this vector will include a constant ($x_{i1} = 1$). The most common way to relate x_i to π is to select one class to serve as a baseline or reference group, and then construct a baseline-category logistic regression model (Agresti, 2002) in which the latent variable L_i is the response. Using the last category $L_i = C$ as the baseline, the model is

$$\begin{aligned} \pi_l(x_i) &= P(L_i = l \mid x_i) \\ &= \frac{\exp(x_i^T \beta_l)}{1 + \sum_{j=1}^{C-1} \exp(x_i^T \beta_j)} \end{aligned} \tag{3.3}$$

for $l = 1, \ldots, C-1$, where $\beta_l = (\beta_{1l}, \beta_{2l}, \ldots, \beta_{ql})^T$ is a $q \times 1$ vector of coefficients of log-odds ratios influencing the probability of belonging to class l relative to the baseline class C. If β_{1l} is positive, the model suggests that a subject with a high level of x_{i1} is relatively more likely to be in class l than a subject with a low value of x_{1i}, adjusting for the other covariates in the model. The summation in the denomination of (3.3) may also be taken from $j = 1$ to $j = C$, if we define $\beta_C = (0, \ldots, 0)^T$ to make the model identifiable.

In this LC regression model, the distribution of L_i is assumed to be affected by the covariates, but the influences of covariates on $Y_{i1}, \ldots, Y_{ip}$ are completely mediated by L_i. That is, conditional on class membership, item responses and the covariates are assumed to be independent,

$$P(Y_{ij} = y_{ij} \mid L_i, x_i) = P(Y_{ij} = y_{ij} \mid L_i).$$

The prevalence of $L_i = l$ is allowed to vary with the covariates, but the meaning of the latent-class variable L_i is still determined only by the items $Y_{i1}, \ldots, Y_{ip}$. The relationship among x_i, L_i and the Y_{ij}'s are as shown in Figure 3.1.

This LC regression model has the attractive property that, if the distribution of Y_i is marginalized over the covariates, it reduces to a standard LC model with the same number of classes C and the same measurement parameters ρ (Bandeen-Roche et al., 1997). Bandeen-Roche et al. suggest that an LC regression model should be constructed by first fitting a standard LC model to Y_i without covariates, which will help the researcher to understand the latent class structure. Covariates may then be added to the model to assess their influence on the latent variable L_i. Introducing covariates in this way does not affect the population parameters ρ, but it may slightly change the estimates of ρ because ρ and β are not precisely orthogonal in the observed-data likelihood function (Chung et al., 2006). The contribution of the ith individual to the observed-data likelihood function can be written as

$$P(Y_i = y_i \mid x_i) = \sum_{l=1}^{C} \pi_l(x_i) \prod_{j=1}^{p} \prod_{m=1}^{M_j} \rho_{jm|l}^{I(y_{ij}=m)}, \tag{3.4}$$

where

$$\pi_l(x_i) = \frac{\exp(x_i^T \beta_l)}{1 + \sum_{j=1}^{C-1} \exp(x_i^T \beta_j)}.$$

In this model, the class membership probabilities $\pi_l(x_i)$ are now conditional probabilities given the covariates, and they are deterministic functions of the β coefficients. If estimates of the marginal class prevalences are desired, they can be obtained

by fitting the model without covariates, or by averaging the estimated values of $\pi_l(x_i)$ over the individuals $i = 1, 2, \ldots, n$ in the sample.

3.2 A Latent-Class Selection Model

Returning now to the missing data problems described in the previous chapters, we will apply the LC regression model to multivariate data with nonignorable missingness. Rather than using the classes to describe the responses to a set of questionnaire items, we will apply the LC model to describe the missingness indicators for these items. Through experience, we have found that information in a set of missingness indicators is sometimes well summarized by a simple latent-class structure, suggesting that a large number of missing-data patterns may be reduced to just a few prototypes. The class membership of any individual is unobserved, but his or her probability of belonging to any particular latent class may be estimated from the multiple missingness indicators. For example, certain types of individuals may tend to say "don't know" for a single item or a group of items. Most of these individuals do not provide usable data for these items. But a few members of that class will answer the questions, and their responses can be used to guess or infer what the missing values for the other subjects might be. Individuals who, based on their probabilities of class membership, look as though they belong to a class, may provide information on the distributions of unseen items within that class.

By adopting a latent-class approach to modeling patterns of missingness, we may avoid the instability and extreme sensitivity of conventional selection models, which posit a direct dependence of missingness for an item on the underlying true value for that item (see, for example, Diggle and Kenward, 1994; Kenward, 1998; Little and Rubin, 2002). Instead of using an incomplete item to predict the probability of missingness for that item, we will use the item to predict class membership, so

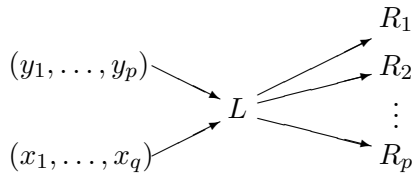

Figure 3.3: Relationship among latent variable, missingness indicators, and items in the LC selection model

that items and missingness are related only through latent classes. Therefore, unlike selection models, our LCSM is expected to be more stable and less sensitive, because the task of predicting a single latent variable L will typically be easier than describing the full distribution of $R_1, \ldots, R_p$. Relationships among covariates, the outcomes, the latent classifier, and the indicators in this proposed model are shown in Figure 3.2.

Latent variable modeling for missing-data mechanisms is not entirely new. Some authors have incorporated latent-class structure into pattern-mixture models to jointly describe the pattern of missingness and the outcome of interest (Lin et al., 2004; Muthen et al., 2003; Roy, 2003). Lin et al. (2004) proposed a latent pattern-mixture model where the mixture patterns are formed from latent classes that link a longitudinal response with a missingness process. Roy (2003) utilized latent classes to model dropouts in longitudinal studies to effectively reduce the number of missing-data patterns. Muthen et al. (2003) also discussed how latent classes could be applied to nonignorable missingness. In each of these models, the latent classes were defined by the missingness indicators and the items themselves. These can be regarded as "shared parameter" models. In our model, which we call a latent-class selection model (LCSM), the classes are defined by the missingness indicators alone. Our classes are simpler to interpret, and the model gives more direct insight into the missing-data process.

In the LCSM, the incomplete items (and perhaps other covariates) are used to

predict class membership as in the LC regression model described in Section 3.1. Our model is essentially the same as LC regression. The only major difference is that now some of the items used as predictors have missing values.

To complete the specification of the LCSM, we will need to apply a parametric model to the joint distribution of the partially missing items. For simplicity, we will assume that these items follow a multivariate normal distribution. The multivariate normal is the most common model applied to the complete data in multivariate missingness problems. Extensions of the LCSM to models for multivariate categorical data, or to mixed data with continuous and categorical variables, are conceptually and computationally straightforward as long as algorithms for handling missing values in these models are available when the missing values are MAR.

Let $y_i = (y_{i1}, y_{i2}, \ldots, y_{ip})$ denote the set of items with nonignorably missing values, and let $x_i = (x_{i1}, x_{i2}, \ldots, x_{iq})$ denote completely observed covariates. We will collect these into a single column vector z_i which contains the y_{ij}'s, the x_{ij}'s, and a constant term. The missingness indicators for y_i will be $r_i = (r_{i1}, r_{i2}, \ldots, r_{ip})$, where $r_{ij} = 1$ if y_{ij} is obseved and $r_{ij} = 0$ if y_{ij} is missing. The missing-data mechanism in the LCSM is

$$\begin{aligned} P(R_i = r_i \mid Z_i = z_i; \beta, \rho) &= \sum_{l=1}^{C} P(R_i = r_i \mid L_i = l) P(L_i = l \mid Z_i = z_i) \\ &= \sum_{l=1}^{C} \pi_l(z_i) \prod_{j=1}^{p} \rho_{j|l}^{r_{ij}} (1 - \rho_{j|l})^{(1-r_{ij})}, \end{aligned} \tag{3.5}$$

where

$$\pi_l(z_i) = \frac{\exp(z_i^T \beta_l)}{1 + \sum_{j=1}^{C-1} \exp(z_i^T \beta_j)},$$

and $\rho_{j|l}$ is the conditional probability that an individual responds to item y_{ij} given that $L_i = l$.

This model (3.5) assumes that the r_{ij}'s are dichotomous, indicating only whether the corresponding y_{ij}'s are observed or missing. In some applications, there may be different types of missing values (e.g., refusal versus "don't know"), and the distribution of the true values underlying the different types of missing values could be different. It may be of interest to distinguish among these types of missing values in the model. Under an MAR assumption, there is no obvious way to do this. Our LCSM, however, immediately extends to missing-data indicators with three or more levels if we replace the model in (3.5) by the more general LC regression model described in Section 3.1. The only change is that the term in (3.5) corresponding to the Bernoulli likelihood for r_{ij} is replaced by a multinomial term for the categories or levels of r_{ij}.

The β and ρ parameters in the LCSM describe how the probabilities of missingness relate to the items and the covariates. These parameters, though interesting, are ultimately a nuisance, because the questions of scientific interest usually pertain to the population of y_i. The parameters of the population distribution of y_i, which we will call θ, are the main target of inquiry, and the β and ρ parameters are important insofar as they affect inferences about θ. In general, y_i will be related to the covariates, and the model for y_i should reflect these relationships. We will therefore write the model for y_i as $P(z_i \mid \theta)$, a joint distribution for y_i and x_i. Alternatively, one could write that model as $P(y_i \mid x_i, \theta)$, as in a multivariate regression of y_i on x_i, because the covariates in x_i have no missing values and therefore do not need to be modeled.

Collecting the missing-data indicators for all subjects into a matrix R, and the y_i's and x_i's into another matrix Z, the likelihood function for this model becomes

$$L(\psi \mid Z, R) \propto \prod_{i=1}^{n} \left[P(z_i \mid \theta) \sum_{l=1}^{C} \pi_l(z_i) \prod_{j=1}^{p} \rho_{j|l}^{r_{ij}} (1 - \rho_{j|l})^{(1-r_{ij})} \right], \tag{3.6}$$

where $\psi = (\theta, \beta, \rho)$ represents all parameters of the population model and the missingness mechanism. The likelihood function (3.6) cannot be used for inference, because

it depends on the missing items in y_i. The likelihood that must be used in practice is

$$L(\psi \mid Y_{obs}, X, R) \propto \prod_{i=1}^{n}\Big[\int P(z_i \mid \theta)\sum_{l=1}^{C}\pi_l(z_i)\prod_{j=1}^{p}\rho_{j|l}^{r_{ij}}(1-\rho_{j|l})^{(1-r_{ij})}\, dY_{mis}\Big], \quad (3.7)$$

where Y_{mis} denotes all the missing items in y_i for all individuals.

Because the likelihood includes a new term $P(z_i \mid \theta)$ and integrates over the missing items in y_i, maximizing this likelihood is more difficult than for the LC regression model in Section 3.1. EM algorithms are no longer straightforward. Rather then attempting to draw inferences about θ based on the likelihood alone, we will apply prior distributions to the parameters and simulate their joint posterior distribution using Markov Chain Monte Carlo (MCMC). In some applications, the ML estimates of some $\rho_{j|l}$'s will be nearly or exactly zero. In those cases, some aspects of β will not be identified (Bandeen-Roche et al., 1997) which will cause difficulty for ML estimation of θ. By applying Bayesian techniques and making the prior distribution for β mildly informative, we will be able to overcome problem of indeterminacy and numerical instability associated with likelihood-based selection modeling.

Chapter 4

Model Fitting Procedures

4.1 MCMC and Multiple Imputation

For our proposed model, an appealing alternative to maximum likelihood estimation is Bayesian inferences based on Markov Chain Monte Carlo (MCMC) (Liu, 2001; Robert and Casella, 2004). By sampling parameters and drawing missing values, MCMC is a natural option for summarizing a posterior distribution without relying on fully-determined density functions or analytical derivatives. In this section, we describe a data augmentation algorithm for multiple imputation (MI) of missing values under our proposed model. Schafer (1997, Chapter 4) points out that, in multivariate missing-data problems, MI can be a convenient alternative to the more common practice of collecting and summarizing a large number of parameter values drawn from their posterior distribution. MI allows an analyst to perform a variety of analyses on the imputed data that do not need to be specified in advance. An introduction to algorithms for Bayesian MCMC and MI in related problems is given by Gelman et al. (2003), Jackman (2000), McLachlan and Krishan (1997), Schafer (1997), Tanner (1996).

To help simplify the notation, we will suppress the dependence on covariates in our probability distributions, so that all distributions will implicitly condition on covariates. In general, MI requires us to obtain independent draws of Y_{mis} from $P(Y_{mis} \mid Y_{obs}, R)$, the posterior predictive distribution of the missing data given the

observed quantities Y_{obs} and R, under a joint model for Y_{com} and R. Under the ignorability assumption, the model for R becomes irrelevant, and Y_{mis} can be drawn from $P(Y_{mis} \mid Y_{obs})$, the posterior predictive distribution ignoring the missing-data mechanism. In our case, the missingness mechanism is nonignorable, so we must generate imputations $Y_{mis}^{(1)}, \ldots, Y_{mis}^{(m)}$ conditioning on the missingness R as well as the observed data Y_{obs}. Once these imputations have been created, further modeling of the missingness becomes unnecessary. The imputed datasets can be analyzed by standard complete-data methods, and information about the missing-data mechanism is carried in the imputed values. For this reason, MI is also an excellent too for sensitivity analyses. Imputed datasets drawn from any joint model for Y_{com} and R may be analyzed in exactly the same way, allowing us to easily compare results across models whose parameterizations may be very different.

Although parameters are not retained in the output of MI, they must still be drawn during the imputation process, because the posterior predictive distribution $P(Y_{mis} \mid Y_{obs}, R)$ incorporates uncertainty about all the parameters of the joint model for Y_{com} and R. In our case, the posterior predictive distribution may be written as

$$P(Y_{mis} \mid Y_{obs}, R) = \int P(Y_{mis} \mid Y_{obs}, R, \psi)\ P(\psi \mid Y_{obs}, R)\ d\psi,$$

where $\psi = (\beta, \rho, \theta)$ are the parameters of the latent-class selection model (LCSM). To obtain the posterior distribution for the parameters, we will need to specify a prior distribution $f(\beta, \rho, \theta)$. Under this prior, the observed-data posterior distribution of the model parameters becomes

$$P(\beta, \rho, \theta \mid Y_{obs}, R) \propto f(\beta, \rho, \theta) \times \prod_{i=1}^{n} \left[\int P(y_i \mid \theta) \sum_{l=1}^{C} \pi_l(y_i) \prod_{j=1}^{p} \prod_{m=1}^{M_j} \rho_{jm|l}^{I(r_{ij}=m)}\ dY_{mis} \right].$$

Because the likelihood function requires integrating out the missing values Y_{mis}, it is difficult to draw from this posterior distribution directly. The computational difficulty would be alleviated, however, if the missing values were known. Thus, it will

be convenient to implement an MCMC procedure using the key idea of data augmentation (Tanner and Wong, 1987), in which missing values and parameters are drawn alternately in sequence.

The MCMC procedure for our LCSM combines the data augmentation procedures for incomplete multivariate data described by Schafer (1997) with an MCMC procedure for LC regression described by Chung et al. (2006). Further simplification will occur if we also treat the latent class indicators $L = (L_1, \ldots, L_n)$ as missing data, filling them in at each cycle, because given Y_{mis} and L, the likelihood for ψ factors into independent likelihoods for θ, β and ρ. If we also choose a prior distribution for ψ in which θ, ρ and β are a priori independent, then the augmented-data posterior for β, ρ, and θ given Y_{obs}, R, Y_{mis}, and L can be partitioned as

$$\begin{aligned}
P(\beta, \rho, \theta \mid Y_{com}, L, R) &\propto f(\rho)f(\beta)f(\theta)P(Y_{com} \mid \theta)P(L, R \mid Y_{com}, \beta, \rho) \\
&\propto f(\rho)f(\beta)f(\theta)\prod_{i=1}^{n} P(y_i \mid \theta)\prod_{l=1}^{C}\left(\pi_l(y_i)\prod_{j=1}^{p}\prod_{m=1}^{M_j}\rho_{jm|l}^{I(r_{ij}=m)}\right)^{I(L_i=l)} \\
&\propto \left[f(\theta)\prod_{i=1}^{n} P(y_i \mid \theta)\right] \\
&\times \left[f(\beta)\prod_{i=1}^{n}\prod_{l=1}^{C}\pi_l(y_i)^{I(L_i=l)}\right] \\
&\times \left[f(\rho)\prod_{l=1}^{C}\prod_{j=1}^{p}\prod_{m=1}^{M_j}\rho_{jm|l}^{n_{jm|l}}\right], \qquad (4.1)
\end{aligned}$$

where x_i and y_i are used as predictors in the LC regression for L_i, and $n_{jm|l} = \sum_{i=1}^{n} I(r_{ij} = m,\ L_i = l)$.

Our MCMC procedure can be regarded as a Gibbs sampler with embedded Metropolis steps to handle parameters for which the conditional distribution cannot be simulated directly. Overall, we will be simulating draws from the joint posterior distribution of Y_{mis}, L, θ, ρ, and β given the observed values of Y_{obs} and R. We will use the superscript "(t)" to denote the simulated value of an unknown quantity at cycle t.

Following the terminology of Tanner and Wong (1987), the algorithm will be divided into two basic steps: an Imputation or I-step, in which the "missing data" (broadly defined) are simulated given assumed values of the parameters, and a Posterior or P-step in which the parameters are drawn from their posterior distribution given assumed values for the missing data. The I-step and P-step are further divided into sub-steps as follows.

1. Imputation step (I-step):
   (a) Draw $Y_{mis}^{(t+1)}$ from $P(Y \mid Y_{obs}, R, L^{(t)}, \theta^{(t)}, \beta^{(t)}, \rho^{(t)})$;
   (b) Draw $L^{(t+1)}$ from $P(L \mid Y_{obs}, R, Y_{mis}^{(t+1)}, \theta^{(t)}, \beta^{(t)}, \rho^{(t)})$.
2. Posterior step (P-step):
   (a) Draw $\theta^{(t+1)}$ from $P(\theta \mid Y_{obs}, R, Y_{mis}^{(t+1)}, L^{(t+1)}, \beta^{(t)}, \rho^{(t)})$;
   (b) Draw $\rho^{(t+1)}$ from $P(\rho \mid Y_{obs}, R, Y_{mis}^{(t+1)}, L^{(t+1)}, \theta^{(t+1)}, \beta^{(t)})$;
   (c) Draw $\beta^{(t+1)}$ from $P(\beta \mid Y_{obs}, R, Y_{mis}^{(t+1)}, L^{(t+1)}, \theta^{(t+1)}, \rho^{(t+1)})$.

In the above algorithm, drawing L, ρ, and θ from their respective conditional posterior distributions is straightforward. The latent classes L_i are drawn from multinomial distributions whose probabilities are obtained by applying Bayes' Theorem to $\pi(y_i)$ and ρ (Chung et al., 2006; Hoijtink, 1998; Garrett and Zeger, 2000; Lanza et al., 2005). Under Dirichlet prior distributions, the posterior distributions of the elements of ρ are also Dirichlet for each item in each latent class. The parameters of the normal complete-data population, θ, are drawn from a standard posterior distribution for the mean vector and covariance matrix. Under the natural priors, this posterior distribution will be a combination of a multivariate normal and inverted Wishart (Schafer, 1997). Simulating β, however, is not trivial because there is no simple conjugate prior family for the coefficients of a multinomial logistic model. The distribution for β is

nonstandard, requiring a Metropolis-Hastings step. Generating Y_{mis}, the missing values in a multivariate normal data matrix is somewhat different from ordinary I-step procedure described by Schafer (1997), because we must consider the information in the latent variables L. Procedures for generating Y_{mis} and β are described in detail below.

Random draws for Y_{mis} are generated from the posterior predictive distribution of Y conditional on L and θ. (Although we are also conditioning on β, ρ, and R, the information in these quantities becomes irrelevant once θ and L are known.) For each subject i, we find the parameters of the normal distribution for the missing Y_{ij}'s given the observed ones. The parameters can be obtained from θ by application of the SWEEP operator (Little and Rubin, 2002; Schafer, 1997). Given these parameters, we draw the missing Y_{ij}'s and then we simulate a value of L_i from its distribution given the now-complete Y_{ij}'s under the LC regression model. If the simulated L_i agrees with the currently assumed class membership of subject i, then the step is complete. Otherwise, we reject the simulated Y_{ij}'s and repeat until agreement in the L_i's is obtained. The computational details of this procedure will be spelled out in Section 5.2.

For the coefficients β of the multinomial logit model, we sample β indirectly from its full conditional distribution using a Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953). At iteration t, a candidate β^* is drawn from a proposal distribution $Q(\beta \mid \beta^{(t)})$ and is compared to the current $\beta^{(t)}$ through ratio of the densities

$$\alpha = min\left(\left[\frac{P(\beta^* \mid Y, L, R)Q(\beta^{(t)} \mid \beta^*)}{P(\beta^{(t)} \mid Y, L, R)Q(\beta^* \mid \beta^{(t)})}\right], 1\right). \tag{4.2}$$

The candidate β^* is promoted to $\beta^{(t+1)}$ with a probability of acceptance α, otherwise we take $\beta^{(t+1)} = \beta^{(t)}$.

A proposal distribution for Metropolis-Hastings should be easy to sample from and should be more diffuse than the target distribution (Gelman et al., 2003). Chung et al. (2006) applied a multivariate Student's t distribution with 4 degrees of freedom, centered at $\beta^{(t)}$ with scale matrix $c^2\Sigma$, where Σ is an estimate of the covariance matrix for β under its full conditional distribution, and c is a constant. Following Gelman et al. (2003), we take $c \approx 2.4/\sqrt{d}$, where d is the number of free parameters in β (in this case, $(C-1)p$). For the guess of Σ, we use the β submatrix of the approximate covariance matrix for the ML estimate of β,

$$\Sigma = -\left(\frac{\partial^2 \ell}{\partial\beta\partial\beta^T}\right)^{-1} |_{\beta=\hat{\beta},\rho=\hat{\rho}},$$

where l is given by

$$l = \sum_{i=1}^{n} \log(P(R_i = r_i)) = \sum_{i=1}^{n}\left(\sum_{l=1}^{C} \pi(y_i) \prod_{j=1}^{p} \prod_{m=1}^{M_j} \rho_{mj|l}^{I(r_{ij}=m)}\right).$$

Starting from initial guesses for the unknown quantities, we repeat the Imputation-Posterior (IP) steps for many iterations to create a sequence of iterates,

$$(Y_{mis}^{(1)}, L^{(1)}, \psi^{(1)}), (Y_{mis}^{(2)}, L^{(2)}, \psi^{(2)}), \ldots, (Y_{mis}^{(t)}, L^{(t)}, \psi^{(t)}), \ldots$$

The stationary distribution to which this sequence converges is the joint posterior distribution of parameters, latent classes, and missing values. The samples of the missing values will converge to the posterior predictive distribution $P(Y_{mis} \mid Y_{obs}, R)$, the distribution from which MI's need to be drawn. After a sufficiently long burn-in period, the simulated missing values—spaced far enough in the sequence to be regarded as independent—can be treated as MI's.

MCMC algorithms for finite-mixture models may exhibit a phenomenon known as label switching, in which the labels for the latent classes suddenly change from one iteration to the next (Celeux et al., 2000). This is a consequence of the fact

that the likelihood function for a finite mixture is invariant to reorderings of the component labels. In our case, we will be summarizing information from the MCMC run only through the imputations of Y_{mis}, or through the simulated values of θ, which is identified apart from the class labels. Even if the class labels permute during the simulation run, the joint posterior distribution of Y_{mis} and θ is unaffected.

Two important practical issues in applying the MCMC procedure are assessing the convergence of the sequence and choosing the number of imputations to be created. The number of iterations needed to achieve stationarity, which is commonly called the burn-in period, should be large enough to eliminate dependence on the starting values. We may view convergence as a lack of serial dependence. If the algorithm has converged by k iterations, then iterates spaced k cycles apart are independent. The value of k is typically chosen by examining the sample autocorrelation function (ACF) plots for one dimensional summaries of the parameters. If autocorrelations for all parameters become negligible by lag k, then a burn-in period of length k should be sufficient. The use of time-series plots, ACF's, and other convergence diagnostics for MCMC are discussed by Ritter and Tanner (1992), Roberts (1992), Liu and Liu (1993), Schafer (1997), and Gelman et al. (2003).

For choosing the number of imputations, proponents of MI often recommend $m = 5$ or $m = 10$ imputations. In typical missing-data problems, 5 or 10 imputations are sufficient to obtain efficient estimates of parameters of the complete-data population (Schafer, 1997). Rubin (1987) showed that the efficiency of an estimate bases on m imputations, relative to an estimate based on an infinite number of them, is approximately

$$\left(1 + \frac{\gamma}{m}\right)^{-1},$$

where γ is the rate of missing information for the quantity being estimated. For example, with 40% missing information, $m = 5$ imputations is 93% efficient, and

$m = 10$ imputations increases the efficiency to 96%. From a standpoint of efficiency alone, there is little incentive to create more than 5 or 10 imputation in a typical problem. There is no harm in taking more, however, and in practice many researchers are now using 25-100 imputations (Graham et al., 2007). Increasing the number of imputation will stabilize p-values and estimated rates of missing information, and it will help eliminate the dependence of results on the choice of an arbitrary random number generator seed.

4.2 Modeling Issues

4.2.1 Choosing a Latent Structure

One of the main modeling issues in using the LCSM is to determine a suitable number of latent classes. The number of classes, C, should be large enough to provide an adequate summary of the missingness. Because of the marginalization property explained in Section 3.1, we may select the number of classes using the same methods as in a standard LC analysis, by fitting conventional LC models (not LC regression) to the missingness indicators R (Bandeen-Roche et al., 1997; Chung et al., 2006). Even without predictors, however, selecting the number of classes is not a simple matter, and there is no universally accepted method for doing so. In practice, researchers who use LC models select the number of classes by examining fit statistics for various models, selecting one that fits the data well and provides an appealing interpretation for the classes.

As the number of classes increases, the fit of an LC model always improves, until the model becomes so large that the parameters are no longer identifiable (Goodman, 1974). However, more classes may increase computational complexity, slow the convergence of MCMC, and make the classes difficult to interpret. In our LCSM, too many classes may destabilize the posterior predictive distribution of Y_{mis}, producing

erratic inferences about the complete-data population. Too few classes, on the other hand, will produce a model that fails to adequately capture the relationships between the complete data Y_{com} and the missingness indicators R. In the most extreme case, a model with $C = 1$ class implies that the missing data are MCAR, producing inferences about the complete-data population that are no different from assuming that nonresponse is ignorable.

A widespread approach to selecting the number of classes is to compare the fit of various models to a saturated model, and increase the number of classes until the fit is judged to be adequate in some absolute sense relative to a saturated model. A likelihood-ratio test of the absolute fit of an LC model is based on the deviance statistic, which compares the actual number of subjects in any response pattern (i.e., with a given set of responses to the items in y_i) to the expected number under the ML estimates for that model. The deviance statistic is

$$G^2 = 2\sum_{j=1}^{J} f_j \log\left(\frac{f_j}{\hat{f}_j}\right),$$

where f_j is the observed frequency for the jth response pattern, and $\hat{f}_j$ is the estimated expected frequency, and J is the total number of possible response patterns. G^2 has a large-sample χ^2 distribution with degrees of freedom equal to J minus the number of estimated parameters. A model with values of G^2 that does not exceed the critical value from the χ^2 table is considered plausible. For the χ^2 approximation to work well, the expected frequencies must be moderately large. A common rule-of-thumb is that at least 80% of the $\hat{f}_i$'s must be at least 5.0, and none should be less than 1.0 (Agresti, 2002). This rule is often violated in LC models, because even if the sample size is large, the observations are often concentrated in a few cells of the contingency table, and other parts of the table are very sparse. When this happens, there is no reliable way to test the absolute fit of the model based on the likelihood function alone.

In ordinary categorical-data modeling, the G^2 statistic may be used to compare non-saturated models of different complexity, because the differences-in-G^2 between two nested models is a likelihood-ratio statistic for testing the simpler model against the more complicated one. In LC analysis, however, the difference-in-G^2 test is not appropriate for comparing models with different numbers of latent classes, because the geometric conditions necessary to obtain a limiting χ^2 distribution is violated (Rubin and Stern, 1994). Likelihood ratio tests pertaining to the number of components in a finite mixture model are non-standard, and limiting distributions for these statistics are mixtures of χ^2 distributions with different degrees of freedom (Lindsay, 1995).

Because of the difficulties associated with G^2 and difference-in-G^2 tests, an increasingly popular way to access the fit of LC models is by posterior predictive checks (Rubin, 1984; Rubin and Stern, 1994; Meng, 1994; Gelman et al., 2005). The posterior predictive check is based on a test statistic that is sensitive to model fit, such as G^2 or the difference in G^2. Rather than comparing the value of this statistic to a frequentist reference distribution, however, it is judged against the posterior distribution that the statistic would have over future datasets from the same population if the model were true. The posterior predictive check distribution (PPCD) uses no large-sample approximations, but simulating it can be computationally intensive. A more detailed description of the PPCD will be given in Section 4.2.2.

Models with different numbers of classes have also been compared by penalized likelihood or information criteria, including the AIC (Akaike, 1987), BIC (Bozdogan, 1987), and CAIC (**?**). Each of these criteria is based on the value of the loglikelihood achieved by the model, adjusted for the number of parameters in the model. The idea behind these criteria is that, given two models that achieve the same loglikelihood, the model with fewer parameters is better. These criteria differ in the penalty applied to each extra parameter. Information on these and related indices is given by Hoijtink

(2001) and on a website maintained by John Uebersax with resources for LC models at http://ourworld.compuserve.com/homepage/jsuebersax. To our knowledge, however, none of these measures has been shown to be effective at selecting the correct number of classes in an LC model. In fact, simulations by Masyn (2003) have shown that all of these criteria perform rather poorly, and should not be relied on when selecting an LC model.

Another way to compare alternative models in a Bayesian setting is through Bayes factors (Berger and Sellke, 1987; Kass and Raftery, 1995). The Bayes factor weighs the alternative models by the posterior evidence in favor of them, comparing the marginal likelihood of two competing models so the model with the largest marginal likelihood is preferred. Bayes factors are difficult to compute, however, and may be sensitive to prior distributions.

4.2.2 Posterior Predictive Checks

Because of the problems associated with the aforementioned approaches, we will rely primarily on the Posterior Predictive Check Distribution (PPCD) of the G^2 statistic, conditional on a particular model and the data, to assess the fit of models with varying numbers of classes. The p-values obtained from PPCD-based tests are honest in the sense that they represent the chance, given the observed data, of seeing a result more unusual than the observed result if the model were correct (Rubin and Stern, 1994).

The PPCD can be described for general problem as follows. Let X denote a sample of data from an arbitrary population. Let $P(X \mid \phi)$ represent a model specification for X, where ϕ represents a parameter with prior distribution $f(\phi)$. If we observe $X = X_{obs}$, then formal inferences about ϕ under this model proceed from the posterior distribution $P(\phi \mid X_{obs})$, which is proportional to $f(\phi) \times P(X_{obs} \mid \phi)$.

The basic idea behind PPCD is to compare features of the observed data with the same features of other datasets that could have been observed under the same model, if the parameters were drawn from $P(\phi \mid X_{obs})$. If replications of the data generated under the model fail to recreate relevant features of the observed data, then the model should be rejected. Let $T(X)$ be any statistic measuring the discrepancy between the model and the data that reveals lack of fit (e.g., the G^2 statistic). We will compare the observed value of this statistic, $T(X_{obs})$, to its PPCD, which is obtained by averaging the distribution $P(T(X) \mid \phi)$ over the posterior distribution $P(\phi \mid X_{obs})$. The PPCD is the posterior predictive distribution of $T(X)$ conditional on the model $P(X \mid \phi)$ and the data X_{obs}. It is the distribution of $T(X)$ that would be expected in replications of the study under the posited model with the same parameters that generated the original data (Gelman et al., 1996). The following steps can be used to simulate the posterior predictive distribution of $T(X)$:

1. Sample $\phi^{(m)}$ from the posterior distribution $P(\phi \mid X_{obs})$.
2. Draw a replicate of the data set $X_{rep}^{(m)}$ from $P(X \mid \phi^{(m)})$.
3. Obtain ML estimates from $X_{rep}^{(m)}$ under the given model and compute $T(X_{rep}^{(m)})$.
4. Repeat steps 1, 2, and 3 for $m = 1, 2, \ldots, M$ to obtain M replicates of $T(X_{rep}^{(m)})$.
5. Compare $T(X_{obs})$ to the empirical distribution of $T(X_{rep}^{(m)})$, $m = 1, 2, \ldots, M$.

If larger values of $T(X)$ indicate worse fit, then the posterior predictive p-value is the tail area probability where $T(X_{rep}^{(m)}) \geq T(X_{obs})$. A small p-value indicates that it is unlikely that the observed data X_{obs} could have come from the posited model. The meaning of the PPCD p-value, and its relationship to frequentist p-values from hypothesis tests, is explored by Meng (1994).

The test mentioned above based on the G^2 statistic will help to assess the absolute fit of a model. To compare models with different numbers of classes, Rubin and Stern (1994) describe a PPCD procedure based on the difference-in-G^2 statistics for the competing models. Suppose we want to test the fit of a C-class model against the fit of a model with $C' > C$ classes. First, we draw a set of model parameter values from their posterior distribution given the data X_{obs} under the C-class model. We then create a simulated dataset of the same size as the original from the C-class model, using the parameter values drawn in the previous step. To this replicate data set, we fit C- and C'-class models and compute the difference-in-G^2 statistics between them. Repeating this procedure many times, we obtain a distribution for the difference-in-G^2 statistics, which becomes a reference distribution to which the observed value of the difference-in-G^2 statistic is compared.

4.2.3 Model for the Complete Data

Implementing MI requires us to specify a population distribution for the complete data. Because MI is a Bayesian procedure, we also need to specify a prior distribution for the parameters of this complete-data model. From this data model and prior distributions, we can obtain a predictive distribution for the missing values conditional on the observed values, which characterizes the relationships between missing values and observed values. Guidelines for choosing a model for the complete data, called an imputation model, are described by Little and Rubin (2002), Schafer (1997), and Schafer and Graham (2002).

In our description of the MCMC procedure in Section 4.1, we supposed that the complete data $Y_{com} = (Y_{obs}, Y_{mis})$ would be described by a multivariate normal distribution. The multivariate normal is the most common imputation model, and most of the software for MI available today is based on this model. The LCSM,

however, is a general concept that can just as easily as be applied to any of the multivariate imputation models described by Schafer (1997). For example, an LCSM for categorical responses can be formulated using loglinear models. Mixed datasets containing both categorical and continuous variables can be described by a general location model, which combines a log-linear model for the categorical variables with a multivariate normal regression for the continuous ones. The LCSM could also be applied to a multivariate linear mixed effects model or a hierarchical linear model for multivariate panel or clustered data (Schafer and Yucel, 2002), which would be appropriate for describing multiple variables collected on a sample of individuals over time, or multiple variables collected on individuals who are grouped together into larger units.

When MI is applied to typical missing-data problems, the specification of the imputation model is not always crucial, because this model is used only to predict the missing parts of the dataset. The robustness of MI to departures from the imputation model are reported from many simulation studies (Ezzati-Rice et al., 1995; Schafer, 1997; Schafer and Graham, 2002). Imputations created under a model that is only a rough approximation to the true population distribution may not have a devastating effect on the final inferences, provided that the analysis method applied to the imputed datasets is reasonable. However, this does not suggest that imputation may be carried out haphazardly. The imputation model should be chosen to be at least approximately compatible with the analysis to be performed on the imputed datasets. In practice, the imputation model should contain at least as much information as the analysis model, and it should be rich enough to preserve the associations or relationships among variables that will be the focus of post-imputation analyses. Relationships between the model used to impute the missing values and the model used to analyze the imputed datasets are explored by Meng (1994) and Schafer (2003).

The main advantage of using a multivariate normal population model is that the computations required are very manageable. With real data, however, multivariate normality rarely holds. Despite natural concerns about non-normality, many researchers have found that a multivariate normal model gives acceptable results even when the variables are binary or ordinal. Imputed values may be rounded to the nearest category, or left unrounded if the analysis procedure allows it. In many settings, the normal model works as well as more complicated alternatives specifically designed for categorical or mixed data. Methods for rounding and the properties of rounded imputations are discussed by Allison (2005) and by Bernaards et al. (2006).

To improve the fit of the normal model, we may also use transformations. A variable that does not appear to be normally distributed may be transformed to approximate normality for purposes of modeling, and the imputed values may be transformed back to the original scale. Box-Cox power transformations and scaled logistic transformations have been implemented in the popular software package NORM (Schafer, 1997). Care must be taken when using power transformations, because if the normal model does not adequately describe the tail behavior of the transformed item, some of the imputed values may be implausibly large or small. With trial and error, a suitable transformation for a variable can usually be found.

4.2.4 Prior Specification

An MCMC procedure requires us to specify prior distributions for all parameters in the model. These prior distributions quantify our beliefs or state of prior knowledge about the parameters. In this section, we discuss how to select priors for parameters of the LCSM.

The functional forms of our prior distributions are chosen primarily for convenience, to make it easy to sample parameters from their respective full-conditional

distributions. As we have seen in Section 4.1, the augmented-data posterior $P(\beta, \rho, \theta \mid Y_{obs}, Y_{mis}, L, R)$ factors into independent posteriors for ρ, β, and θ if the priors on these parameters are independent. When the priors are independent, the full-conditional distributions for these parameters given Y_{mis} and L may be expressed as

$$P(\theta \mid Y_{com}, L, R) \propto f(\theta) \prod_{i=1}^{n} P(Y_i \mid \theta), \tag{4.3}$$

$$P(\beta \mid Y_{com}, L, R) \propto f(\beta) \prod_{i=1}^{n} \prod_{l=1}^{C} \pi_l(y_i)^{I(L_i=l)}, \tag{4.4}$$

$$P(\rho \mid Y_{com}, L, R) \propto f(\rho) \prod_{l=1}^{C} \prod_{j=1}^{p} \prod_{m=1}^{M_j} \rho_{jm|l}^{n_{jm|l}}. \tag{4.5}$$

In Equation (4.3), the functional form of the full-conditional distribution for θ depends on a specific imputation model adopted by the imputer. The problem of choosing a prior for θ is no different under the LCSM than under an assumption of ignorable nonresponse. In practice, it is usually acceptable to place a noninformative prior on θ, as described by Schafer (1997). The priors for β and ρ, however, are specific to the LCSM and must be chosen carefully.

The class membership probabilities $\pi_l(y_i)$ in the LCSM are determined by β, the coefficients of the baseline-category multinomial logit model. With little available prior information, any Dirichlet with identical values for hyperparameter α_l, $l = 1, \ldots, C$, between 0 and 1 can be assigned to π directly as a noninformative prior. This includes Dirichlet with $\alpha = (1, \ldots, 1)$ (i.e., uniform prior) and Dirichlet with $\alpha = (1/2, \ldots, 1/2)$ (i.e., Jeffreys' prior). Another way way to avoid placing a prior for β is to create fictitious fractional observations and spread them in a judicious way across the covariate patterns, the unique values of Y_i and x_i appearing in the dataset. Priors of this type, which are data-dependent, are discussed by Clogg and Eliason (1987) and Clogg et al. (1991) to stabilize the estimates in sparse tables. This type of prior would be attractive and convenient if all the predictors in the logit model

were completely observed. In the LCSM, however, the predictors include summaries of the items which are occasionally missing.

If we instead create an explicit prior for β, a natural choice is to vectorize β and apply a multivariate normal distribution on the β coefficents. Multivariate normal priors for the coefficients of a logistic model are discussed by Dellaportas and Smith (1993) and by Knuiman and Speed (1988). Ridge regression can be viewed as a Bayesian technique with an exchangeable normal prior distribution on the coefficients (Goldstein, 1976). For LC regression, Chung et al. (2006) suggest a product of p-dimensional multivariate normal distributions for β, which can be considered as an analogue to a ridge prior for the multinomial logit model. If we allow the inverses of the prior covariance matrices to approach zero, we obtain an improper uniform prior distribution for β. A uniform prior can sometimes perform well for the LCSM. When the probabilities of missingness on some items within some classes are close to one, however, some elements of β may become nearly unidentified. When running data augmentation with unidentified parameters, these non-identified parameters may drift to extremely large values and produce numeric overflow. To prevent this from happening, we may add a small amount of prior information to stabilize the estimated coefficients. The details and implications of various choices of priors are described by Heinze and Schemper (2002) and Galindo-Garre et al. (2004).

For the LC measurement parameters ρ, it is natural to apply independent conjugate prior distribution to the response probabilities for each item in each class. Letting $\rho_{j|l}$ denote the vector of response probabilities for item Y_{ij} in class $L_i = l$, the Dirichlet density is

$$P(\rho_{j|l}) \propto \prod_{m=1}^{M_j} \rho_{jm|l}^{\alpha_m - 1}, \tag{4.6}$$

where the α's are user-specified positive hyperparameters. The resulting full-conditional

posterior for ρ would have the same form as the full-conditional likelihood, with prior observations added to each class. To reflect a state of prior ignorance, we could set the hyperparameters equal to 1/2 (producing a Jeffreys' prior) or to 1 (producing a uniform prior). In practice, the difference between these priors tends to have little or no observable impact on the results from the LCSM unless the sample size is very small.

The practice of selecting prior distributions for the LCSM is best demonstrated by example. We will show how to apply prior distributions in the next chapter, when we use the LCSM on the ALEXSA dataset. Whenever possible, it makes sense to try a variety of alternative prior distributions to see how they affect the results. In the applications we have tried, changes in the results under different reasonable priors are barely noticeable.

Chapter 5

Application

5.1 Foreign Language Attitude Scale

5.1.1 Description of Data

Before examining the ALEXSA dataset mentioned in Chapter 1, we use a simpler data example to demonstrate how a large number of missingness patterns can often be adequately summarized by a small number of latent classes. The data summarized in Table 5.1, collected by Raymond (1987) and analyzed by Schafer (1997), were obtained from undergraduates enrolled in foreign language courses at The Pennsylvania State University in the early 1980's (Raymond and Roberts, 1983). Twelve variables were collected on a sample of $n = 279$ to investigate the usefulness of a newly developed instrument, the Foreign Language Attitude Scale (FLAS), for predicting success in the study of foreign language. Descriptions of the variables, along with the number of missing values for each one, are provided in Table 5.1.

In this example, six variables (LAN, AGE, PRI, SEX, FLAS, HGPA) are excluded from modeling the missingness mechanism, because those variables have few or no missing values. However, these variables do enter as covariates into the regression part of the LCSM, because the values of these items may influence missingness on other items. Restricting attention to the six variables with substantial missingness (MLAT, SATV, SATM, ENG, CGPA, GRD), a summary of the missingness patterns and the number of observations in each pattern is given in Table 5.2.

Table 5.1: Variables in Foreign Language Achievement Study

Variable	*Description*	*Missing*
LAN	foreign language studied	0
	(1=French, 2=Spanish, 3=German, 4=Russian)	
AGE	age group (1=less than 20, 2=21+)	11
PRI	number of prior foreign language courses	11
	(1=none, 2=1-2, 3=3+)	
SEX	1=male, 2=female	1
FLAS	score on FLAS	0
MLAT	Modern Language Aptitude Test	49
SATV	Scholastic Aptitude Test, verbal	34
SATM	Scholastic Aptitude Test, math	34
ENG	score on PSU English placement exam	37
HGPA	high school grade point average	1
CGPA	current college grade point average	34
GRD	final grade in foreign language course	47
	(2=A, 1=B or lower)	

Table 5.2: Missingness patterns of the FLAS data (1=observed, 0=missing)

frequency	*MLAT*	*SATV*	*SATM*	*ENG*	*CGPA*	*GRD*
177	1	1	1	1	1	1
29	0	1	1	1	1	1
1	1	1	1	0	1	1
23	1	0	0	0	0	1
2	0	0	0	0	0	1
20	1	1	1	1	1	0
16	0	1	1	1	1	0
2	1	1	1	0	1	0
7	1	0	0	0	0	0
2	0	0	0	0	0	0

5.1.2 Fitting a Latent-Class Model

The first step in applying the LCSM is to fit an LC model to the patterns shown in Table 5.2 and select the number of classes. As we described in Section 4.2, our strategy is to fit models without covariates at first, starting with 2 classes and proceeding to three, four, and so on, evaluating the fit of each model using various criteria. Deviance statistics for two-, three-, and four-class LC models are shown in Table 5.3. No p-values are given in this table, because the χ^2 approximation is not appropriate. Comparing differences in G^2 to χ^2 distribution is not appropriate either, for the reasons mentioned in Chapter 4.

Nevertheless, this table suggest that three classes fits substantially better than two, and four classes is essentially no better than three. Based on this table alone, it is tempting to use a three-class model. When we examined the estimated ρ-parameters

Table 5.3: Deviance Statistics for LC models applied to FLAS data

Description	G^2	*DF*
2 Latent Classes	22.01	50
3 Latent Classes	3.73	43
4 Latent Classes	2.57	36

Table 5.4: Estimated probabilities of responding to each item under the two-class model, and estimated class prevalences

Missingness indicator	*Class I*	*Class II*
MLAT	0.882	0.816
SATV	0.000	1.000
SATM	0.000	1.000
ENG	0.000	0.988
CGPA	0.000	1.000
GRD	0.735	0.854
Prevalence	*0.122*	*0.878*

for the three-class model, however, we found that two of the classes were similar in their tendencies to response to the six items, so we decided to use a two-class model.

Estimates of the ρ-parameters for the two-class model are shown in Table 5.4. The values in this table are estimated probabilities of responding to each item within each class. From this table, we see that a large majority of participants (estimated at 88%) were likely to respond to each item. The remaining participants (estimated at 12%) had high probabilities of providing MLAT and GRD, but gave no data from SATV, SATM, ENG, or CGPA. The missingness patterns in this dataset can thus be

described as follows: SATV, SATM, ENG, and CGPA were missing together for about 12% of the study participants, and missing values for MLAT and GRD were essentially random.

5.1.3 Analysis by Multiple Imputation

Schafer (1997, Chapter 6) analyzed this dataset by multiply imputing the missing values under an assumption of MAR. He replaced the nominal variable LAN with three dummy indicators to distinguish among the four language groups and applied a multivariate normal imputation model to the resulting 14 variables. We will also assume that the complete data, Y_{com}, for the 14 variables is distributed as a multivariate normal with μ and Σ, but we will describe the missingness indicators R by a two-class LCSM. Letting y_i denote the 14 items and r_i the vector of missingness indicators, we assume that

$$\begin{aligned} y_i &\sim N_p(\mu, \Sigma), \\ P(r_i \mid y_i) &= \sum_{l=1}^{2} \pi_l(y_i) \prod_{j=1}^{6} \rho_{j|l}^{r_{ij}} (1-\rho_{j|l})^{(1-r_{ij})}, \end{aligned}$$

where

$$\pi_l(y_i) = \frac{\exp(y_i^T \beta_l)}{1 + \exp(y_i^T \beta_1)}$$

for $l = 1$.

Schafer (1997) discovered that a standard noninformative prior could not be applied to μ and Σ, because all of the values of GRD happened to be missing for students in the Russian language group (LAN=4), which causes the partial correlations between GRD and the language dummy variables to be inestimable. Following Schafer (1997), we centered and scaled the observed values for each variable to have mean 0 and variance 1, and applied a mildly informative prior distribution to Σ analogous to the kind of prior used in ridge regression. This prior effectively smooths the correlations

in Σ toward zero, with information equivalent to a prior sample size of three. For β_1, the vector of logistic coefficients, we applied the multivariate normal prior distribution

$$\beta_1 \sim N_p \left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 5 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix} \right], \tag{5.1}$$

which smooths the logistic coefficients toward zero. For ρ, we applied the Dirichlet prior distribution with hyperparameters 1/2 (i.e., Jeffreys' prior).

Applying the MCMC algorithm described in Chapter 4, we created $m = 20$ multiple imputations of the missing data under the LCSM by running 20 independent chains for 1000 steps each. After imputation, we analyzed the imputed datasets by the same method used by Schafer (1997), fitting a logistic regression model to each imputed dataset to predict the probability of $GRD = 2$. Covariates in the logistic model included three dummy indicators for language, dummy indicators for AGE=2 and SEX=2, linear and quadratic contrasts for PRI ($\text{PRI}_L = -1, 0, 1$ and $\text{PRI}_Q = 1, -2, 1$ for PRI=1,2,3), and main effects for FLAS, MLAT, SATV, SATM, ENG, HGPA and CGPA.

ML estimates and standard errors for the logistic coefficients were computed for each of the 20 imputed datasets, and the 20 sets of results were then combined using Rubin's rules (1987) for multiple-imputation inference with scalar estimands. Table 5.5 summarizes these results, including the point estimate $\overline{Q}$, the standard error $\sqrt{T}$, and the t-statistic $\overline{Q}/\sqrt{T}$. For comparison, this table shows results from the LCSM and from a multiply-imputed analysis that assumes the missing values are missing at random (NORM). Very little difference is seen in any of the quantities, except the standard error for the coefficient of LAN_4. This is the effect that cannot be estimated from the observed data, for which essentially all information is provided by the prior

distribution for Σ.

The two methods, LCSM and NORM, apply identical models to the complete data population and identical prior distributions to the parameters of the complete-data model. The methods differ only in their assumptions about missing values. NORM assumes the missing values are MAR, whereas LCSM describes the missingness as a mixture of two MCAR mechanisms. In the first mechanism, four variables are missing with very high probability, and in the other mechanism, all variables are observed with high probability. Intuitively, LCSM is probabilistically assigning each subject to one mechanism or the other, and imputing the missing items for each subject given his or her observed items, using an imputation model based on the other subjects belonging to that mechanism. NORM, which assumes MAR, imputes the missing items for each subject from a single imputation model for all subjects. NORM treats all subjects identically regardless of their missingness pattern, whereas LCSM groups subjects with similar missingness patterns.

In the two-class LCSM for this example, the values of four variables (SATV, SATM, ENG, and CGPA) are essentially always missing for subjects in Class 1. That class provides essentially no information on how to impute the missing values for those four variables. Because of the mildly informative prior for β_1, which smooths the logistic coefficients in the LCSM toward zero, the LCSM allows some information to flow from Class 2 to Class 1 for imputing those missing values. When adequate information is present within a class, the LCSM takes advantage of that class-specific information, but otherwise the LCSM borrows information across the classes in a manner similar to an MAR procedure. In this example, therefore, it is not surprising that LCSM and NORM produced similar results. In the next example, however, we will see a noticeable difference between the two methods.

Table 5.5: MI inferences for logistic regression coefficients under the two-class LCSM and an assumption of ignorable nonresponse (NORM)

	$\overline{Q}$		$\sqrt{T}$		$\overline{Q}/\sqrt{T}$	
	LCSM	NORM	LCSM	NORM	LCSM	NORM
Intercept	-15.1	-15.5	2.95	3.07	-5.10	-5.07
LAN_2	0.361	0.312	0.511	0.518	0.71	0.60
LAN_3	1.15	1.12	0.436	0.453	2.64	2.48
LAN_4	-0.357	-0.110	1.51	4.13	-0.24	-0.03
AGE_2	1.41	1.40	0.455	0.457	3.09	3.07
PRI_L	0.263	0.350	0.254	0.261	1.04	1.34
PRI_Q	-0.115	-0.165	0.145	0.150	-0.79	-1.10
SEX_2	0.797	0.861	0.451	0.443	1.77	1.94
FLAS	0.0382	0.0386	0.016	0.0166	2.39	2.33
MLAT	0.110	0.114	0.0481	0.0480	2.28	2.37
SATV	-0.0038	-0.0033	0.0033	0.0033	-1.15	-1.01
SATM	0.0002	-0.0004	0.0028	0.0026	0.06	0.13
ENG	0.0107	0.0110	0.0237	0.0238	0.45	0.46
HGPA	2.16	2.27	0.438	0.439	4.93	5.1
CGPA	0.911	0.809	0.539	0.588	1.69	1.38

5.2 The ALEXSA Data

5.2.1 Description

Over the last four decades, behavioral scientists and health researchers have studied predictors of addiction and other problem behaviors (APB's). Preventive interventions are often designed to alter these predictors (e.g., dysregulation) in hopes of subsequently reducing APB's (Kusche and Greenberg, 1994; Tarter et al., 2002). A critical step in prevention research is learning which APB predictors are "elevated" in an at-risk community to target in an intervention. The ALEXSA instrument was designed to measure correlates of APB's in children and young adolescents. The data we are examining came from the first wave of a two-wave pilot study to validate the instrument by measuring the test-retest reliabilities of the 76 self-report items. Our analysis will focus on a section that measures levels of social and environmental exposures to alcohol. Responses to these items are combined into an alcohol risk index, which is thought to be predictive of future alcohol use and other negative outcomes (e.g., conduct disorder).

At the beginning of the alcohol section, participants were shown a set of pictures (a beer can, a glass and bottle of wine, a shot glass, a liquor bottle, and a martini) and were asked to identify the substance that was depicted in all of the illustration. If the participant responded with "alcohol", "beer", "wine" or a similar answer, he or she was queried with additional items related to alcohol risk. If the participant responded incorrectly or said "I don't know" the participant skipped all the remaining items on alcohol risk and proceeded to the next section. In some cases, the child correctly identified the substance, but the data collector mis-typed the response when entering it into the computer, inadvertently causing the remaining items to be skipped when they should not have been. Among those who correctly identified alcohol and proceeded, some said "I don't know" or refused to answer one or more of the remaining items.

Table 5.6: Alcohol-related variables in the ALEXSA pilot study

Variable	Description	Range
RECOG	Do you know what it is in the set of pictures?	0, 1
DEA.1	How do you first learn about alcohol?	1, . . . , 8
	(1=parent 2=sibling 3=grandparent/other relative 4=friend	
	5=teacher/other school leader 6=media 7=religious leader 8=other)	
DEA.2	Do you know anyone who drinks alcohol? (0=No 1=Yes)	0, 1
DEA.3	How do your parents feel about kids drinking alcohol?	0, 1, 2, 3
	(↑ means more positive attitude)	
DEA.4	How do most of your friends feel about kids drinking?	0, 1, 2, 3
	(↑ means more positive attitude)	
DEA.5	How do you feel about kids your age drinking alcohol?	0, 1, 2, 3
	(↑ means more positive attitude)	
DEA.6	Has anyone ever offered you a drink of alcohol? (0=No 1=Yes)	0, 1
DEA.7	Have you ever drank alcohol, even just a sip? (0=No 1=Yes)	0, 1
DEA.8	How difficult would it be for you to get some?	0, 1, 2, 3
	(↑ means more easiness)	
DEA.9	Will you drink alcohol when you are in high school?	0, 1, 2, 3
	(↑ means more frequency)	
DEA.10	How often do you drink alcohol right now?	0, 1, 2, 3
	(↑ means more frequency)	

Table 5.7: Frequency and types of missing values for alcohol items in ALEXSA pilot study with $n = 319$ participants (DK=don't know, RF=refused)

Variable	Missing	Missing Types
RECOG	101	DK=92, Misspecified=9
DEA.1	121	RF=6, DK=14, Skipped=101
DEA.2	110	RF=4, DK=5, Skipped=101
DEA.3	113	RF=5, DK=7, Skipped=101
DEA.4	122	RF=3, DK=18, Skipped=101
DEA.5	109	RF=4, DK=4, Skipped=101
DEA.6	115	RF=10, DK=4, Skipped=101
DEA.7	114	RF=8, DK=5, Skipped=101
DEA.8	115	RF=9, DK=5, Skipped=101
DEA.9	118	RF=5, DK=12, Skipped=101
DEA.10	105	RF=4, DK=0, Skipped=101

The missing values in the resulting dataset, therefore, are a mixture of legitimately skipped questions, inadvertent skips, don't-know responses, and refusals.

The items used in this analysis are listed in Table 5.6, and the number and type of missing values for each item are shown in Table 5.7. Each of the items after the stem question (RECOG) is binary or ordinal. The alcohol risk index is computed as a weighted sum of those items, and is defined only for those who recognized the drug. Among the $n = 319$ participants in this study, 206 correctly identified alcohol and proceeded to the remaining items; 92 participants answered "I don't know" to the stem question; and 12 participants mistakenly identified alcohol as other drugs; and 9 participants were mistakenly coded as not knowing what alcohol is, even though they apparently did, causing the remaining items to be skipped inadvertently.

The results from this study will be heavily influenced by how we handle the data from the stem question. In an ideal world, the stem question would allow us to unambiguously divide the participants into two groups: those who recognize alcohol, and those who do not. The $206 + 9 = 215$ participants who answered alcohol, beer, wine, etc. are presumably cognizant enough to answer the remaining questions. It also seems reasonable to assume that the 12 who mistakenly identified alcohol as something else do not have sufficient knowledge to provide trustworthy answers to DEA.1—DEA10. The crucial issue is how to handle the 92 participants who said "I don't know." One possibility is to assume that none of these individuals knew about alcohol and that they could not knowledgeably answer the remaining questions. It is difficult to believe, however, that nearly one third of children in this study had no awareness of the substance. Rather, the researchers believed that in many cases, a response of "I don't know" simply indicated that the participant was uncooperative and did not want to provide information at that time. The most reasonable way to proceed, in our opinion, is to regard these as a mixture of individuals who are truly

naive and individuals who really do know about alcohol. Additional covariates—especially age—may help us to estimate the probability for any child of belonging to one group or the other.

In the analysis, we apply an LCSM to multiply impute the missing values of the alcohol items and compare the results to those from an assumption that nonresponse is ignorable. We will interpret the stem question as having 215 positive responses (indicating sufficient awareness of alcohol), 12 negative responses (indicating insufficient awareness), and 92 missing values. For the latter, we will impute a binary response (positive or negative) as we impute responses to the missing values for DEA.2–DEA.10. If the imputed value for the stem question is positive, we will aggregate the imputed values for DEA.2–DEA.10 for that individual into an overall score for the alcohol risk index; if the imputed value for the stem question is negative, we will ignore the imputations for DEA.2–DEA.10 and leave the alcohol risk index undefined.

Covariates will play a crucial role in this imputation procedure, because they will help us to distinguish those who know about alcohol from those who do not. Covariates used in this analysis, which are listed in Table 5.8, include sex, age, number of siblings (which is thought to influence exposure), and an ordinal measure of academic achievement. The SCHOOL variable describes the environment from which the child was recruited. The $n = 319$ participants in this study were drawn from three places. Some (SCHOOL=1) attended an ordinary public school. Others (SCHOOL=2) attended an enrichment program for children with learning disabilities and other special needs. The remaining children (SCHOOL=3) were housed in an inpatient facility where they were receiving treatment for various psychiatric conditions. The researchers drew participants from these three diverse environments because they wanted to demonstrate the utility of ALEXSA to measure predictors of APB's across

Table 5.8: Descriptions of covariates from the ALEXSA data

Variable	*Description*
SEX	0=female, 1=male
AGE	8, 9, 10, 11, 12, 13
SIB	Total number of siblings (0, . . . , 10)
GRADE	A=5, B=4, C=3, D=2, F=1
SCHOOL	Ordinary=1, Enhancement=2, Psychiatric=3
CON.1	Broke curfew 3+ times? (0=no, 1=yes)
CON.2	Lied to get of doing something 3+ times (0=no, 1=yes)
CON.3	Ever skipped school? (0=no, 1=yes)
CON.4	Stolen worth $10+ from family? (0=no, 1=yes)
CON.5	Stolen worth $10+ from non-family? (0=no, 1=yes)
CON.6	Ever started a fight? (0=no, 1=yes)
CON.7	Ever scare someone to get way? (0=no, 1=yes)
CON.8	Breaking & Entering (0=no, 1=yes)
CON.9	Break into car (0=no, 1=yes)
CON.10	Ever vandalized? (0=no, 1=yes)
CON.11	Ever start fire where not supposed to be? (0=no, 1=yes)

populations with diverse risk. The remaining covariates in Table 5.8 come from another section of ALEXSA pertaining to delinquency and related problem behaviors. These items will be aggregated into an overall index of conduct disorder. One of our post-imputation analyses will involve the correlation between the alcohol risk and conduct disorder indices, which helps us to validate both of these measures. The conduct disorder items contain a small number of missing values, but missingness on this section was much less problematic than for the alcohol portion. Missing values for the conduct disorder items will be imputed, but their missingness indicators will not be modeled.

5.2.2 Identifying a latent-class structure

As with the FLAS data, we begin this analysis by fitting various LC models to the missingness indicators to identify an appropriate number of classes. As we pointed out in Chapter 3, the LCSM does not need to regard the missingness indicators as binary. If different types of missing values are present, we can distinguish among them to obtain a richer description of the missing-data mechanism. Consequently, we will model the missingness indicators for DEA.1–DEA.10 as nominal with four levels: 1=observed, 2=skipped, 3=refused, and 4=don't know. In addition, we will include the stem question in the LC model, regarding it as nominal with three levels: 1=correctly indentified alcohol, 2=don't know, and 3=incorrectly identified alcohol as something else. This item is not purely a missingness indicator, but a combination of missing-data codes and responses regarded as known. Because the stem question plays such an important role in this analysis, we coded the item this way in the latent-class portion of the LCSM to provide the richest possible description of the classes. Later, we will code the stem question as binary (know alcohol, does not know alcohol), introduce it as a predictor in the LC regression part, and impute values for the "don't

Table 5.9: Fit statistics for latent-class models describing the missingness indicators in the ALEXSA alcohol data

Number of classes	G^2	*# of par*
2	418.323	33
3	271.809	66
4	208.320	99
5	181.614	122

know" responses that will represent the subject's true underlying state of knowledge about the substance.

We began by fitting an LC model with two classes using an EM algorithm, and then increased the number of classes to three, four, and five. The deviance fit statistics and number of parameters for each of these models are shown in Table 5.9. As before, we do not report any likelihood-based p-values for these G^2 statistics, because the data are too sparse for the usual χ^2 approximations to be reliable. Rather than attempting to select a model by likelihood values alone, we will assess the fit of these models by comparing the G^2 statistics to their PPCD's.

To simulate the PPCD's, we need to specify prior distributions for the parameters of the traditional LC model. We applied independent Dirichlet prior distributions to the class prevalences and item-response probability within each class, setting the hyperparameters α to 1/2, which corresponds to the Jeffreys' noninformative prior. We then simulated 1,000 random draws of G^2 from each model using the procedure of Rubin and Stern (1994). The posterior predictive p-values, defined as the proportion of simulated G^2 values exceeding the observed G^2 for each model, are shown in Table 5.10. From these results, it appears that three-, four-, and five-class models all

Table 5.10: Posterior predictive p-values for latent-class models for ALEXSA alcohol data

Model Description	G^2	*p-value*
2 Latent Classes	418.3226	0.000
3 Latent Classes	271.8089	0.701
4 Latent Classes	208.3202	0.997
5 Latent Classes	181.6136	0.999

describe the data fairly well, at least with respect to this one index of overall fit.

In addition to examining the overall fit of each model, it is also helpful to compare the fit of one model to another. The difference-in-G^2 statistics for these nested models should not be compared to χ^2 distributions, because here the regularity conditions needed to obtain a limiting χ^2 distribution fail (Aitkin and Rubin, 1985; Ghosh and Sen, 1985; Lindsay, 1995). Therefore, we will compare each difference-in-G^2 statistic (ΔG^2) to simulated values drawn from its PPCD distribution. Using the procedure described by Rubin and Stern (1994), we drew 1000 values each from the PPCD's of ΔG^2 to compare 3 versus 4, 3 versus 5, and 4 versus 5 classes. For comparing 3 versus 4 classes, and 3 versus 5 classes, essentially all 1,000 simulated values fall below the actual ΔG^2, indicating that the 3-class model could be rejected in favor of the 4-class and 5-class models. For comparing the 4-class to the 5-class model, however, 893 of the 1,000 simulated ΔG^2 values exceeded the actual statistic, indicating that there is little evidence to prefer 5 classes to 4 classes.

Based on these PPCD's, we decided that the missingness mechanism is best described by four classes. Maximum-likelihood estimates of the class prevalences and measurement parameters from the four-class model are reported in Table 5.12. The

Table 5.11: Posterior predictive p-values for nested model comparisons

# Classes Null	*# Classes Alt.*	$\triangle G^2$	*p-value*
3	4	63.49	0.001
3	5	90.20	0.001
4	5	27.61	0.893

values of the measurement parameters suggest that Class 1 represents those who correctly identify alcohol and answer all the remaining questions with high probability. Class 2 represents those who tend to respond "I don't know" or provide incorrect answers to the stem question, causing the data collector to skip the remaining items. About 8% of the individuals in Class 2, however, do apparently know what alcohol is, so this class also includes those for whom the remaining items were incorrectly skipped. Class 3 represents those who were moderately likely to refuse to answer DEA.1–DEA.10, and Class 4 represents those who were moderately likely to say "I don't know". We estimate that about 55% of the subjects belong to Class 1, 35% belong to Class, 3% belong to Class 3, and 6% belong to Class 4. The 45 response patterns in the raw data have now been succinctly summarized by just four prototypes.

Notice that in Class 3, the probability of refusing to answer individual items is moderately high, but some persons in this class do provide answers to each item. This is good news for the LCSM, because it shows that the observed data contain useful information for imputing the missing values in this class. Similarly, in Class 4, the rates of "don't know" are elevated relative to the other classes, but a majority within the class still answers any item, so we have a firm basis to impute the missing values within this class as well. The most problematic group is Class 2, whose members were told to skip items DEA.1–DEA.10 with probability 1. For the children in this class

Table 5.12: ML estimates under the four-class model

Items	Class 1 (55.3 %)				Class 2 (35.4 %)			
	Cor	DK	Incor		Cor	DK	Incor	
RECOG	1.00	0.00	0.00		0.08	0.81	0.11	
	Ans	Skip	Refuse	DK	Ans	Skip	Refuse	DK
DEA.1	0.96	0.00	0.00	0.03	0.00	1.00	0.00	0.00
DEA.2	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
DEA.3	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
DEA.4	0.96	0.00	0.00	0.04	0.00	1.00	0.00	0.00
DEA.5	0.99	0.00	0.00	0.01	0.00	1.00	0.00	0.00
DEA.6	0.97	0.00	0.03	0.00	0.00	1.00	0.00	0.00
DEA.7	0.97	0.00	0.03	0.00	0.00	1.00	0.00	0.00
DEA.8	0.98	0.00	0.01	0.01	0.00	1.00	0.00	0.00
DEA.9	0.96	0.00	0.01	0.04	0.00	1.00	0.00	0.00
DEA.10	0.99	0.00	0.01	0.00	0.00	1.00	0.00	0.00
Items	Class 3 (2.8 %)				Class 4 (6.4 %)			
	Cor	DK	Incor		Cor	DK	Incor	
RECOG	1.00	0.00	0.00		1.00	0.00	0.00	
	Ans	Skip	Refuse	DK	Ans	Skip	Refuse	DK
DEA.1	0.45	0.00	0.44	0.11	0.57	0.00	0.06	0.38
DEA.2	0.79	0.00	0.21	0.00	0.69	0.00	0.10	0.21
DEA.3	0.56	0.00	0.44	0.00	0.63	0.00	0.05	0.32
DEA.4	0.67	0.00	0.33	0.00	0.49	0.00	0.00	0.51
DEA.5	0.56	0.00	0.44	0.00	0.90	0.00	0.00	0.10
DEA.6	0.55	0.00	0.45	0.00	0.80	0.00	0.00	0.20
DEA.7	0.66	0.00	0.34	0.00	0.75	0.00	0.00	0.25
DEA.8	0.24	0.00	0.76	0.00	0.81	0.00	0.05	0.14
DEA.9	0.45	0.00	0.43	0.11	0.80	0.00	0.00	0.20
DEA.10	0.67	0.00	0.33	0.00	1.00	0.00	0.00	0.00

who know what alcohol is, the missing values for the remaining items will have to be imputed by gleaning information from the other classes. The LCSM will have to borrow responses from children in other classes whose covariate values are similar to theirs. The LCSM will also need to predict the true state of alcohol knowledge for the children in the class who responded "I don't know" to the stem question. These predictions will also based on information gleaned from the other covariates in the model.

5.2.3 Specifying the remaining parts of the LCSM

The latent-class structure describing the missing-data indicators is quite interesting, but ultimately it is only a device to help us predict the missing values of the alcohol items for each individual. To finish the specification of the LCSM, we will need to apply a population model to the items themselves, and regress the latent classes on the items.

Following the common practice of researchers who use MI, we will apply a multivariate normal population model to the items in this dataset, treating the binary and ordinal items as if they are continuous for purposes of imputation. We are primarily interested in the overall alcohol risk index, which is a weighted sum of responses to DEA.2–DEA.10. We want to characterize the distribution of this index, and preserve its basic relationships to the other covariates. Therefore, we will apply a multivariate normal distribution to DEA.2–DEA.10 and all the covariates listed in Table 5.8. The nominal variable DEA.1 is omitted from this model because it is not needed for the risk index. The nominal SCHOOL, which has three levels, is replaced by two dummy indicators to distinguish among the three levels. The stem question, RECOG, will enter this model as a binary indicator for whether the child can recognize alcohol. Finally, all of the variables in this normal model will enter the multinomial logistic

regression as main effects to predict the probabilities of membership in Classes 1–4.

5.2.4 Prior distributions

With a multivariate normal distribution for the complete data, it is customary to apply improper noninformative prior distributions to μ and Σ. In this example, however, we cannot use an improper prior for Σ, because of the presence of the stem question. Among those who do not recognize alcohol, responses to the remaining alcohol items are never observed because they are undefined. For these subjects, imputed values for DEA.2–DEA.10 are irrelevant and will never be used. Nevertheless, the partial correlations between the stem question and these items are inestimable and will cause problems during the MCMC run. To overcome these problems, we employ a mild informative ridge prior distribution as described by Schafer (1997, Chapter 5). We suppose that, given Σ, μ is conditionally multivariate normal, and that Σ is inverted-Wishart,

$$\mu \mid \Sigma \sim N(\mu_0, \tau^{-1}\Sigma),$$

and

$$\Sigma \sim W^{-1}(m, \Lambda),$$

where $\tau > 0, \Lambda > 0$ and m are user-specified hyperparameters. Then the augmented-data posterior distribution as $\tau \to 0$ for fixed m and Λ is given by

$$\mu \mid \Sigma, Z \sim N(\bar{z}, \frac{1}{n}\Sigma),$$

and

$$\Sigma \mid Z \sim W^{-1}[n + m, (\Lambda^{-1} + nS)^{-1}],$$

where $\bar{z}$ is the sample mean vector of Z and S is the sample covariance matrix of Z (with a denominator of n). With this posterior distribution, the covariances Σ has been smoothed toward a matrix proportional to Λ^{-1} (Schafer, 1997, Chapter

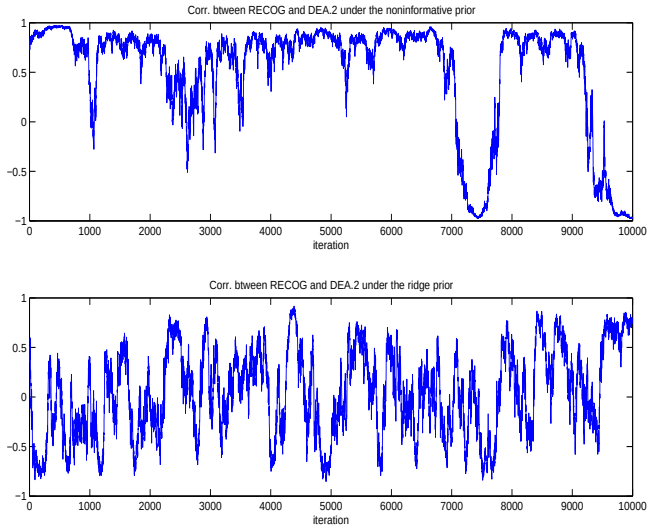

Figure 5.1: Time-series plots of correlation between the stem question and DEA.2 over 10000 iterations of data augmentation under the usual noninformative prior and the ridge prior

5). The degree of smoothing is controlled by a single hyperparameter which can be interpreted as a number of prior observations of (Y_{com}, X). Even a small positive value of this hyperparameter—say, 1% of the sample size or less—will usually be sufficient to stabilize estimation. After centering and scaling the observed data for each variable to have mean 0 and variance 1, we applied this prior with $\tau = 0$, $m = \epsilon$, and $\Lambda^{-1} = \epsilon I_d$ for $\epsilon = 3$, where I_d is the $d \times d$ identity matrix. This prior effectively smooths the correlation matrix for Z slightly toward the identity matrix, supplying information equivalent to that of three fictitious subjects. Under this prior, the augmented-data posterior becomes

$$\mu \mid \Sigma, Z \sim N(\bar{z}, n^{-1}\Sigma), \tag{5.2}$$

$$\Sigma \mid Z \sim W^{-1}(n+3, [3I_d + nS]^{-1}), \tag{5.3}$$

where $\bar{z} = n^{-1}\sum_{i=1}^{n} z_i$ and $S = n^{-1}\sum_{i=1}^{n}(z_i - \hat{z})(z_i - \hat{z})^T$.

To see how this type of ridge prior fixes inestimability problem, we performed a

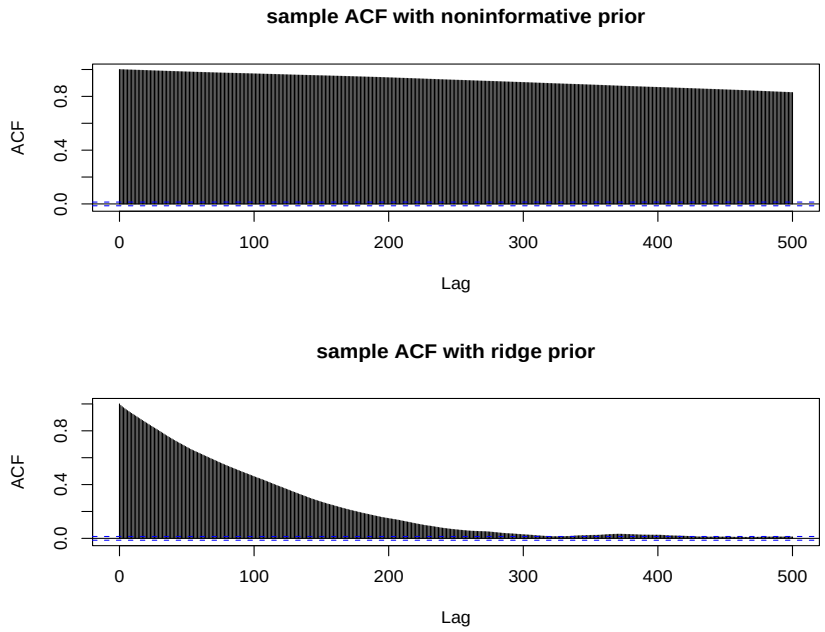

Figure 5.2: Sample ACFs for correlation between the stem question and DEA.2 over 10,000 iterations of data augmentation under the noninformative prior and the ridge prior

long run of data augmentation under the noninformative prior and the ridge prior and constructed time-series plots for correlation between the stem question and DEA.2 for the purpose of demonstration. Time-series plots for this estimate over 10,000 iterations of data augmentation are provided in Figure 5.1. The iterates of estimate under the noninformative prior wander to the boundary of the parameter space, whereas the series for this estimate under the ridge prior appear to approach stationarity even though they show long-range dependence. Sample ACFs for the correlation between the stem question and DEA.2 under both priors are displayed in Figure 5.2. Under the ridge prior, serial correlations are slowly converging but died out by lag 500.

For the measurement parameters ρ, we applied a Jeffreys' prior independently to the vectors of ρ's for each item within each class. For the coefficients β of the multinomial logit model, we applied mildly informative normal prior distributions to

the coefficient vectors $\beta_1, \beta_2, \beta_3$,

$$\beta_1, \ldots, \beta_3 \sim \prod_{l=1}^{3} N_p \left[\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 5 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix} \right]. \tag{5.4}$$

5.2.5 Results: Alcohol risk and exposure

Using these prior distributions, we ran the MCMC procedure of Chapter 4 as a single chain. Starting values for μ and Σ were computed from the incomplete y_i's under an assumption of MAR, using the EM algorithm from multivariate normal data described by Schafer (1997, Chapter 5). Results from the first 10,000 cycles were discarded, and the chain was run for an additional 50,000 cycles.

In ordinary circumstances, we would analyze the results from the LCSM using multiple imputation. That is, we would save the simulated values of Y_{mis} at every kth cycle after burn-in, where k is suitably large, and treat these imputations as independent draws of Y_{mis} from $P(Y_{mis} \mid Y_{obs}, R)$. In this example, however, an interesting question arises over how to handle the imputed responses to the stem question. In the y_i vector, this item is coded as 0 and 1, corresponding to not recognizing and recognizing alcohol, respectively, and the subsequent alcohol items are regarded as relevant only when the recognition item is 1. Because we are approximating the joint distribution of the items by a multivariate normal model, however, the imputed values for the recognition question are continuously distributed. In practice, we would advise an analyst to round off these imputed values to 0 or 1 using a cutoff 1/2 or the adaptive cutoff rule of Bernaards et al. (2006), and then consider the remaining alcohol items only for the individuals for whom the rounded value of recognition is 1. However, Allison (2005) and Horton et al. (2003) point out that rounding can sometimes intro-

duce bias, and they recommend that imputed values be left unrounded whenever the subsequent analysis allows it.

In this example, we are mainly interested in the overall alcohol risk index, an weighted sum of DEA.2–DEA.10, for those who recognize alcohol, and we want to see what the LCSM implies about this measure apart from any biases due to rounding. Therefore, rather than working with rounded imputed data, we will instead directly examine the simulated parameters from the 50,000 cycles of MCMC after burn-in. These can be regarded as a dependent, serially correlated sample from the posterior distribution of μ and Σ given Y_{obs} and R. Using well known properties of the multivariate normal distribution, we can compute from any simulated draw of (μ, Σ) the implied conditional mean value of the alcohol risk index given that the subject correctly recognizes alcohol (RECOG=1). By applying this same transformation to each simulated draw of (μ, Σ), we obtain 50,000 dependent draws of this parameter from its posterior distribution given Y_{obs} and R. The average of these 50,000 draws is a Monte Carlo estimate of the posterior mean for this parameter, and their standard deviation is an estimate of the posterior standard deviation, a Bayesian version of the standard error (SE).

Exposure to alcohol varies by age and sex, and because these variables are in the imputation model, we can also transform any draw of (μ, Σ) to obtain the implied mean alcohol risk index among those who recognize alcohol within any category of age and sex. Estimates of these means and SE's, computed from the 50,000 draws of parameters generated by the LCSM, are reported for boys and girls of each age in Table 5.13. For comparison, this table also shows the estimates and SE's under a multivariate normal population model assuming that the missing values are missing at random (MAR). In every category of age and sex, the estimate from the LCSM is substantially higher than the estimate that assumes MAR. The estimates from this

Table 5.13: Mean estimate and standard error (SE) of alcohol risk index for subjects who correctly identified alcohol, by categories of age and sex, under the latent-class selection model (LCSM) and under an assumption of missing at random (MAR)

	AGE	8	9	10	11	12	13
LCSM	Girl	2.0937	3.1368	4.1799	5.2229	6.2660	7.3091
	(SE)	(1.0238)	(0.7639)	(0.5832)	(0.5638)	(0.7189)	(0.9680)
	Boy	4.0171	5.0602	6.1033	7.1464	8.1895	9.2325
	(SE)	(1.0200)	(0.7527)	(0.5602)	(0.5313)	(0.6870)	(0.9396)
MAR	Girl	1.4566	2.5694	3.6822	4.7950	5.9079	7.0207
	(SE)	(0.9503)	(0.7049)	(0.5342)	(0.5185)	(0.6686)	(0.9056)
	Boy	3.5020	4.6148	5.7276	6.8405	7.9533	9.0661
	(SE)	(0.9508)	(0.6984)	(0.5161)	(0.4896)	(0.6387)	(0.8781)

table are also plotted for boys and girls in Figure 5.3. The differences between the lines for LCSM and MAR represent a substantively meaningful discrepancy that could easily impact statistical inferences about a population from a sample of this size. A rule-of-thumb that we often use is that, if the bias in a parameter estimate exceeds 40%–50% of its standard error, then that bias will adversely impact the performance of confidence intervals and tests. The differences between many of the estimates from LCSM and MAR exceed 50% of their standard errors, showing that the discrepancies are inferentially relevant.

To understand why the LCSM and MAR models give different results for this example, we examined the imputed values for missing items in several of the simulated draws of Y_{mis}. We rounded the continuously distributed imputed values for RECOG to 0 or 1, using a cutoff of 1/2. In the datasets imputed under the MAR assumption, nearly all of the 92 subjects with missing values for the stem question were assigned

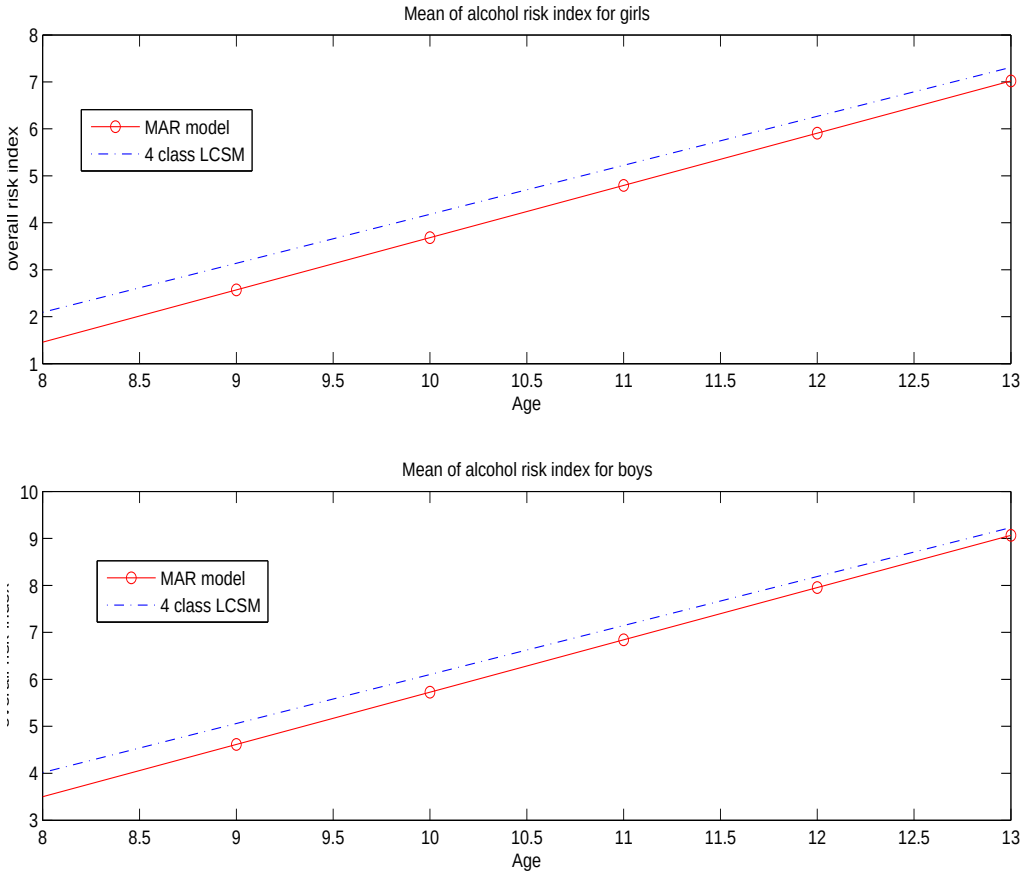

Figure 5.3: Estimates of mean alcohol index for those who recognize alcohol by age and sex under the latent-class selection model (LCSM) and under an assumption of missing at random (MAR)

Table 5.14: Mean estimate and standard error (SE) of the correlation between the alcohol risk and conduct disorder indices for subjects who correctly identified alcohol under the latent-class selection model (LCSM) and under an assumption of missing at random (MAR)

Corr	Model	Estimate	Variance
ρ	LCSM	0.4612	0.0030
	MAR	0.4865	0.0028
ρ_{sex}	LCSM	0.4405	0.0032
	MAR	0.4648	0.0029
ρ_{age}	LCSM	0.4293	0.0033
	MAR	0.4524	0.0031
$\rho_{age \times sex}$	LCSM	0.4295	0.0033
	MAR	0.4535	0.0030

to RECOG=1; only one or two were assigned to RECOG=0. In the datasets imputed under the LCSM, approximately 73 out of the 92 subjects were assigned to RECOG=1, and about 19 were assigned to RECOG=0. In effect, the MAR model estimates that almost every subject who provides an ambiguous answer to the stem question really does recognize alcohol, whereas the LCSM estimates that about 75–80% of these subjects recognize alcohol. The subjects who provide these ambiguous answers tend to be younger than average, and their covariates suggest that they have below-average levels of alcohol exposure. By assigning more of these low-exposure children to RECOG=0, the LCSM effectively removes them from the pool for which the risk index is computed, raising the overall mean risk.

To see if there is any potential impact on relationships to other indices (e.g., conduct disorder) under different assumptions about missingness, we calculated cor-

relations between the alcohol risk and conduct disorder indices. The conduct disorder index is defined as the average of all 11 items listed in Table 5.8. Simple and partial correlations are provided in Table 5.14. In Table 5.14, ρ_{sex}, ρ_{age}, and $\rho_{age\times sex}$ represent partial correlations that measure the degree of association between alcohol risk and conduct disorder, with the effect of sex, age, and age×sex removed respectively. Results from the two models are similar, suggesting there is little impact of differing assumptions about missingness.

In this example, the data provide no empirical information to favor one model over the other. The fact that the MAR assumption implies that nearly every child who does not definitely answer the stem question actually does recognize alcohol, however, is very telling. This result, in our opinion, provides *prima facie* evidence that the missingness mechanism is not MAR. We do not know whether the LCSM is correct. In fact, we are certain that the LCSM is wrong, as all statistical models are. Nevertheless, we have shown that a four-class LC model provides a reasonable summary of the complicated missing data patterns in this example, and when these four classes are related to the missing items in a simple fashion, the imputed values for the missing items look more reasonable than they do under MAR.

From a researcher's perspective, this example underscores the need to carefully consider how to design data-collection procedures when answers to some questions will determine whether or not other questions are regarded as relevant. Measurement error in a stem question can propagate to other questions, introducing potentially large biases. The procedures used in the ALEXSA pilot study directed the data collector to skip all alcohol-related items if the participant gave an incorrect or ambiguous answer to the alcohol recognition item. That procedure would make sense if none of these participants knew what alcohol was. Both of the models we examined (LCSM and MAR), however, strongly suggested that a majority of those who responded "I don't

know" to the stem question really did know about alcohol. Graham et al. (2006) has recently argued that, in a situation like this, it would be wise to collect responses to the subsequent items for at least a random subsample of those who did not correctly identify alcohol. Moving from deterministic skip rules to probabilistic skip rules will produce datasets from which crucial assumptions about these stem questions can be formally tested, leading to richer and more plausible models for the incomplete data.

Chapter 6

Discussion

6.1 Conclusions

An incomplete multivariate dataset does not contain information that allows us to identify the nonignorable aspects of the missingness mechanism unless we make other unverifiable assumptions. Until now, researchers have explored the implications of MNAR missing data by fitting selection models and pattern-mixture models. These models, even if they make sense in longitudinal studies with dropout, are poorly suited to multivariate datasets with complicated missingness patterns. In this thesis, we have argued that a nonignorable model based on the idea of latent classes of response behavior, is a natural and useful alternative to a standard MAR analysis.

We believe that this new approach may avoid much of the extreme sensitivity to untestable distributional assumptions. We also believe that the new model will be more attractive to social and behavioral scientists than conventional pattern-mixture models, because the latent classes provide an intuitively appealing description of the types of respondents found in the study.

Given the unknown true nature of missing data—whether it is MAR or MNAR—we view this new method primarily as a tool for sensitivity analysis. When researchers cannot be sure about the distributional form or correctness of a model, then the most responsible and objective way to proceed is to present alternative results from a variety of plausible models. In practice, missing values will arise for many reasons, and miss-

ing values are not all alike. Ambiguous, refused or skipped items are not necessarily the same, and it seems useful to have models that can distinguish among these various types of missingness, allowing them to have different relationships with the outcomes of interest. The ALEXSA application in Chapter 5 shows that different assumptions about missing data can indeed lead to different results, with important substantive implications. Because suspicions of nonignorable missingness are quite common, we expect that the LCSM will be of interest to researchers in many fields.

6.2 Future work

In this thesis, we have proceeded under the assumption of a multivariate normal distribution for the complete data. The normal model is a natural starting point for this method, but it is also quite limiting. The multivariate normal model implies linear additive relationships among the variables which may be too simplistic. Rounding of imputed values, at best, only a rough approximation to more plausible models for binary and ordinal responses. In the future, we will be extending our method to more principled models for multivariate discrete responses. One promising alternative is a multivariate probit model which characterizes binary and ordinal responses as coarsened versions of continuously distributed noraml variates. These models, which can be regarded as a natural extension of the multivariate normal, will be able to handle larger numbers of variables than the log-linear approaches described by Schafer (1997, Chapter 7–9).

To better understand our method's properties, it is desirable to perform simulations to compare this method to MAR and MNAR alternatives under a variety of missing-data mechanisms. Simulations of this type are rarely done in practice. Some might regard them as artificial, because in every realistic example the true mechanism is unknown. Nevertheless, it would be interesting to see whether this model performs

any better or worse than other methods when its assumptions are violated. In a few rare instances, some or all of the missing values later become known. For example, results from the survey regarding the Slovenian plebiscite analyzed by Rubin et al. (1995) could be compared to actual results from the population poll that took place one month later. The missing-data patterns in that example were too simple to be summarized by an LCSM. If we are able to find a richer multivariate example where some of the missing values are available later, it would provide an interesting and useful test for this model and its competitors. For a practical simulation setting, we may follow Collins et al. (2001)'s scheme which compared the performance of MI and ML under different strategies and missing-data mechanisms.

To make these methods accessible to broader communities of researchers, we eventually want to develop software that implements the Bayesian MCMC procedure for the LCSM in a reliable and user-friendly-manner. The LCSM procedure can be implemented as an add-on to existing procedures for multiple imputation, but it will require some additional considerations. The user must be able to try different number of latent classes and compare the fit of alternative LC models using criteria such as PPCD's. Automatic procedures for eliciting prior distributions and recommendations for these priors will also be needed.

Bibliography

Agresti, A. (2002). *Categorical Data Analysis, Second edition.* J. Wiley & Sons, New York.

Aitkin, M. and Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, Series B*, 47:67–75.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, 52:317–332.

Allison, P. D. (2005). Imputation of categorical variables with PROC MI. SAS Users Group International conference, Philadelphia, PA.

Baker, S. G., Ko, C. W., and Graubard, B. S. (2003). A sensitivity analysis for non-randomly missing categorical data arising from a national health disability survey. *Biostatistics*, 4:41–56.

Baker, S. G. and Laird, N. M. (1988). Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *Journal of American Statistical Association*, 83:62–69.

Bandeen-Roche, K., Miglioretti, D. L., Zegar, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92:1375–1386.

Bartholomew, D. and Knott, M. (1999). *Latent variable models and factor analysis, volume 7 of kendall's library of statistics.* Oxford Universtiy Press.

Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society, Series B*, 37:129–145.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: the ireconciability of p values and evidence. *Journal of the American Statistical Association*, 82:112–122.

Bernaards, C. A., Belin, T. R., and Schafer, J. L. (2006). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Statistics in Medicine*, 26:1368–1382.

Bozdogan, H. (1987). Model-selection and Akaike's information criterion (AIC): The genenral theory and its analytical extensions. *Psychometrika*, 52:345–370.

Celeux, G., Hurn, M., and Robert, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95:957–970.

Chambers, R. L. and Welsh, A. H. (1993). Log-linear models for survey data with nonignorable non-response. *Journal of the Royal Statistical Society, Series B*, 55:157–170.

Chung, H., Flaherty, B. P., and Schafer, J. L. (2006). Latent class logistic regression: application to marijuana use and attitudes among high school seniors. *Journal of the Royal Statistical Society, Series A*, 169:723–743.

Clogg, C. C. (1995). *Latent class models. In Handbook of statistical modeling for the social and behaviral sciences (eds. G. Arminger, C. C. Clogg, and M. E. Sobel).* New York: plenum.

Clogg, C. C. and Eliason, S. R. (1987). Some common problems in log-linear analysis. *Sociological Methods and Research*, 16:8–44.

Clogg, C. C. and Goodman, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association*, 79:762–771.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using bayesian logistic regression. *Journal of the American Statistical Association*, 86:68–78.

Collins, L. M., Schafer, J. L., and Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing-data procedures. *Psycological Methods*, 6:330–351.

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

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

Daniels, M. and Hogan, J. (2000). Reparameterizing the pattern mixture model for sensitivity analysis under informative dropout in longitudinal studies. *Biometrics*, 56:1241–1249.

Dayton, C. M. and Macready, G. B. (1988). Concomitant-variable latent class models. *Journal of the American Statistical Association*, 83:173–178.

Dellaportas, P. and Smith, A. F. M. (1993). Bayesian inference for generalized linear and proportional hazards models via gibbs sampling. *Applied Statistics*, 42:443–460.

Demirtas, H. and Schafer, J. L. (2003). On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 22:2553–2575.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Diggle, P. and Kenward, M. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics*, 43:49–73.

Ezzati-Rice, T. M., Johnson, W., Khare, M., Little, R. J. A., Rubin, D. B., and Schafer, J. L. (1995). A simulation study to evaluate the performance of model-based multiple imputations in nchs health examination surveys. *Proceedings of the Annual Research Conference*, pages 257–266, Bureau of the Census, Washington, DC.

Fairclough, D. L., Peterson, H. F., Cella, D., and Bonomi, P. (1998). Comparison of several model-based methods for analysing incomplete quality of life data in cancer clinical trials. *Statistics in Medicine*, 17:781–796.

Fitzmaurice, G. M., Cliord, P., and Heath, A. F. (1996a). Logistic retression models for binary panel data with attrition. *Journal of the Royal Statistical Society, Series A*, 59:249–263.

Fitzmaurice, G. M., Laird, N. M., and Zahner, G. E. P. (1996b). Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, 91:99–108.

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

Forster, J. J. and Smith, P. W. F. (1998). Model-based inference for categorical survey data subject to nonignorable nonresponse (with discussion). *Journal of the Royal Statistical Society, Series B*, 60:57–70.

Galindo-Garre, F., Vermunt, J. K., and Bergsma, W. P. (2004). Bayesian posterior estimation of logit classification. *Sociological Methods and Research*, 33:88–117.

Garrett, E. S. and Zeger, S. L. (2000). Latent class models. *Biometrics*, 56:1055–1067.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis, Second edition.* Chapman and Hall, London.

Gelman, A., Meng, X. L., and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statistical Sinica*, 6:733–807.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. F., and Meulders, M. (2005). Multiple imputation for model checking: completed-data plots wiht missing and latent data. *Biometrics*, 61:74–85.

Ghosh, K. and Sen, S. K. (1985). On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results. *In Proceedings of the Berkeley conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II. (eds. L. M. Le cam, and R. A. Olshen)*, pages 789–806.

Glynn, R. J., Laird, N. M., and Rubin, D. B. (1986). *Selection modelling versus mixture modelling with nonignorable nonresponse. In Drawing Inference from Self-selected Samples (ed. H. Wainer).* Springer: New York.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61:215–231.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. Prevention Science.

Graham, J. W., Palen, L., and Olchowski, A. E. (2006). Students aren't perfect: Data quality, and other methodological issues in the healthwise south africa projects. The joint prevention and methodology seminar.

Greene, W. (1991). *LIMDEP User's manual, Version 6.0.* New York: Econometric Software, Inc.

Haberman, S. J. (1979). *Aanlysis of qualitative data, volume 2.* New York: Academic Press.

Hagenaars, J. A. (1990). *Categorical longitudinal data-Loglinear analysis of panel, trend and cohort data.* Sage, Newbury.

Hagenaars, J. A. (1993). *Loglinear models with latent variables.* Sage Publications.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and its applications. *Biometrika*, 57:97–109.

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

Heddeker, D. and Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2:64–78.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21:2409–2419.

Heitjan, D. F. and Landis, L. R. (1994). Assessing secular trends in blood pressure a multiple-imputation approach. *Journal of the American Statistical Association*, 89:750–759.

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

Hogan, J. W., Roy, J., and Korkontzelou, C. (2004). Tutorial in biostatistics: handling drop-out in longitudinal studies. *Statistics in Medicine*, 23:1455–1497.

Hoijtink, H. (1998). Constrained latent class analysis using the gibbs sampler and posterior predictive p-values: Applications to educational testing. *Statistical Sinica*, 8:691–711.

Hoijtink, H. (2001). Confirmatory latent class analysis: Model selection using bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research*, 16:717–735.

Horton, N. J., Lipsitz, S. P., and Parzen, M. (2003). A potential for bias when rounding in multiple imputation. *The American Statistician*, 57:229–232.

Ibrahim, J., Chen, M., Lipsitz, S., and Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100:332–346.

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

Jackman, S. (2000). Estimation and inference via bayesian simulation: An introduction to markov chain monte carlo. *American Journal of Political Science*, 44:375–404.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Statistics in Medicine*, 17:2723–2732.

Kenward, M. G. and Molenberghs, G. (1999). Parametric models for incomplete continuous and categorical longitudinal data. *Statistical Methods in Medical Research*, 8:51–83.

Knuiman, M. K. and Speed, T. P. (1988). Incorporating prior information into the analysis of contingency tables. *Biometrics*, 44:1061–1071.

Kusche, C. A. and Greenberg, M. T. (1994). The PATHS curriculum. Seattle, WA: Developmental Research and Programs.

Lanza, S. T., Collins, L. M., Schafer, J. L., and Flaherty, B. P. (2005). Using data augmentation to obtain standard errors and conduct hypothesis tests in latent class and latent transition analysis. *Phychological Methods*, 10:84–100.

Lazarsfeld, P. E. and Henry, N. W. (1968). *Latent Structure Analysis.* Boston: Houghton Mifflin.

Lillard, L. and Panis, C. (2000). *aML user's guide and reference manual.* Los Angeles: EconoWare.

Lin, H., McCulloch, C. E., and Rosenheck, R. A. (2004). Latent pattern-mixture models for informative intermittent missing data in longitudinal studies. *Biometrics*, 60:295–305.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications.* Institute of Mathematical Statistics.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data, Second edition.* John wiley & Sons, New York.

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

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

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

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

Liu, C. and Liu, J. (1993). Disscussion on the meeting on the gibbs sampler amd other markov chain monte carlo methods. *Journal of the Royal Statistical Society, Series B*, 55:82–83.

Liu, J. S. (2001). *Monte carlo strategies in scientific computing.* New York: springer-Verlag.

Masyn, K. E. (2003). What's new in latent class enumeration for general growth mixture models. Spring meetings of the International Biometric Sociery Eastern North American Region (ENAR), Tampa, FL.

McCutcheon, A. L. (1987). *Latent Class Analysis.* Sage University Paper series on Quantitative Applications in the Social Sciences, 07-064. Sage Publications: Beverly Hills, CA.

McLachlan, G. J. and Krishan, T. (1997). *The EM Algorithm and Extension.* J. Wiley & Sons, New York.

Meng, X. L. (1994). Posterior predictive p-values. *Annals of Statistics*, 22:1142–1160.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091.

Michiels, B., Molenberghs, G., and Lipsitz, S. R. (1999). Selection models and pattern-mixture models for incomplete data with covariates. *Biometrics*, 55:978–983.

Molenberghs, G., Kenward, M. G., and Goetghebeur, E. (2001). Sensitivity analysis for incomplete contingency talbes: the slovenian plebiscite case. *Journal of the Royal Statistical Society, Series C*, 50:15–29.

Molenberghs, G., Kenward, M. G., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, 84:33–44.

Muthen, B. O., Jo, B., and Brown, C. H. (2003). Comment on the barnard, hill, & rubin article, principal stratification approach to broken randomized experiments: A case study of school choice vouchers in new york city. *Journal of the Amercian Statistical Association*, 98:311–314.

Park, T. (1998). An approach to categorical data with nonignorable nonresponse. *Biometrics*, 54:1579–1590.

Pirie, P. L., Murray, D. M., and Leupker, R. V. (1988). Smoking prevalence in a cohort of adolescents, including absentees, dropouts, and transfers. *American Journal of Public Health*, 78:176–178.

Raymond, M. R. (1987). An introductive approach to analyzing incomplete multi-varaite data. Presented at the annual meeting of the American Educational Research Association, April 20–24, 1987, Washington, DC.

Raymond, M. R. and Roberts, D. M. (1983). Development and validation of a foreign language attitude scale. *Educational and Psycological Measurement*, 43:1239–1246.

Ridenour, T. A., Clark, D. B., Greenberg, M. T., Minners, S., Singer, L. T., Tarter, R. E., and Cottler, L. B. Reliability and validity of the assessment of liability and exposure to substance use and antisocial behavior (ALEXSA) in 9 to 12 year old students. Under review.

Ritter, c. and Tanner, M. A. (1992). Facilitating the gibbs sampler: The gibbs stopper and the griddy-gibbs sampler. *Journal of the American Statistical Association*, 87:861–868.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods, Second edition.* Springer-Verlag, New York.

Roberts, G. O. (1992). *Convergence diagnostics of the Gibbs sampler. In Bayesian Statistics 4 (eds. J. M. Bernardo, J. Berger, A. P. Dawid, and A. F. M. Smith).* Oxford: Oxford University Press.

Robins, J. M. and Gill, R. D. (1997). Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine*, 16:39–56.

Roy, J. (2003). Modeling longitudinal data with non-ignorable dropouts using a latent dropout class model. *Biometrics*, 59:829–836.

Rubin, D. B. (1974). Characterizing the estimation of parameters in incomplete data problems. *Journal of the American Statistical Association*, 69:467–474.

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

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 4:1151–1172.

Rubin, D. B. (1987). *Multiple imputation for survey nonresponse.* J. Wiley & Sons, New York.

Rubin, D. B., Stern, H., and Vehovar, V. (1995). Handling 'Don't know' survey responses: The case of the slovenian plebiscite. *Journal of the American Statistical Association*, 90:822–828.

Rubin, D. B. and Stern, H. S. (1994). *Testing in Latent class models using a posterior predictive check distribution. In Latent variable analysis: Applications for developmental research (eds. A. Von Eye and C. C. Clogg).* Thousand Oaks, CA: Sage.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman and Hall, New York.

Schafer, J. L. (2003). Multiple imputation in multivariate problems where the imputer's and analyst's models differ. *Statistica Neerlandica*, 57:19–35.

Schafer, J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7:147–177.

Schafer, J. L. and Yucel, R. (2002). Computational strategies for multivariate linear mixed-effect models with missing values. *Journal of Computational and Graphical Statistics*, 11:421–442.

Scheid, S. A selection model for bivariate normal data, with a flexible nonparamteric missing model and a focus on variance estimates. Under review.

Smith, D. M., Robertson, W. H., and Diggle, P. J. (1996). Oswald: Object-oriented software for the analysis of longitudinal data in S. Technical Report MA 96/192, Department of Mathematics and Statistics, University of Lancaster, LA14Yf, U. K.

Stasny, E. A. (1987). Some markov-chain models for nonresponse in estimating gross labor force flows. *Journal of Official Statistics*, 3:359–373.

Stasny, E. A. (1988). Modeling nonignorable nonresponse in categorical panel data with an example in estimating gross labor flows. *Journal of Business and Exonomic Statistics*, 6:207–219.

Stasny, E. A. (1990). Symmetry in flows among reported victimization classifications with nonresponse. *Survey Methodology*, 2:305–330.

Tanner, M. A. (1996). *Tools for statistical inference, Third edition, Springer Series in Statistics*. Springer, New York.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82:528–550.

Tarter, R. E., Kirisci, L., Vanyukov, M., Cornelius, J., Pajer, K., Shoal, G., and Giancola, P. (2002). Predicting adolescent violence: Impact of family history, substance use, psychiastric history, and social adjustment. *American Journal of Psychiatry*, 159:1541–1547.

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

Titterington, D. A., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, New York.

Troxel, A. B., Harrington, D. P., and Lipsitz, S. R. (1998). Analysis of longitudinal measurements with nonignorable non-monotone missing values. *Applied Statistics*, 47:425–438.

Uebersax, J. S. and Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine*, 9:559–572.

Vach, W. and Blettner, M. (1995). Logistic regression with incompletely observed categorical covariates - investigating the sensitivity against violation of the missing at random assumption. *Statistics in Medicine*, 14:1315–1330.

Verbeke, G. and Molenberghs, G. (2000). *Statistical Methods in Medical Research*. Springer-Verlag, New York.

Verbeke, G., Mollenbergh, G., Thijs, H., Lesaffre, E., and Kenward, M. G. (2001). Sensitivity analysis for nonrandom dropout: A local influence approach. *Biometrics*, 57:7–14.

Wu, M. C. and Bailey, K. R. (1984). Estimation and comparison of changes in the presence of informative right censoring: Conditional linear model. *Biometrics*, 45:939–955.

Wu, M. C. and Bailey, K. R. (1988). Analysing changes in the presence of informative right censoring caused by death and withdrawl. *Statistics in Medicine*, 7:337–346.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presense of informative right censoring by modeling the censoring process. *Biometrics*, 48:765–779.

Vita

Hyekyung Jung was born in Kwangju, Korea, on October 15, 1974. In February 1997 she graduated Magna Cum Laude from Seoul National University with a B.A. degree in Consumer Studies and Resource Management. Later that year she was admitted to the M.A. program in Consumer Studies and Resource Management at Seoul National University and received a M.A. degree in February 1999. During the following two years, she was employed as a full-time assistant at Research Institute of Human Ecology in Seoul National University. In August 2001 she entered The Pennsylvania State University and obtained the Doctor of Philosophy Degree in Statistics in August 2007.