Ignorable and Nonignorable Modeling in Regression with Incomplete Covariates

by
Nanhua Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2011

Doctoral Committee:
Professor Roderick J. Little, Chair
Professor Susan A. Murphy
Professor Trivellore E. Raghunathan
Associate Professor Michael R. Elliott

To Yuanshu and my parents

AKNOWLEDGEMENTS

I owe my deepest gratitude to my advisor, Dr. Roderick J. Little, whose guidance, encouragement, and support from the preliminary to the concluding level enabled me to develop an understanding of the subject. His ideas have helped shape the framework of my dissertation. I am grateful to Drs. Michael R. Elliott, Trivellore E. Raghunathan, and Susan A. Murphy for their input and help in the accomplishment of this dissertation.

Personally, I would like to thank my inspiring parents and lovely wife, for their endless love and support. It is through them I realized the value of my professional career.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Missing data is an important practical problem in many applications of statistics. We consider multivariate regression with missing data. Reviews of previous research on the topic include Little (1993), Ibrahim et al. (1999), Ibrahim et al. (2002), Ibrahim et al. (2005), and Chen et al. (2008). Three approaches are:

(a) Complete-case analysis (CC), which discards the incomplete cases;

(b) Ignorable likelihood (IL) methods, which base inferences on the observed likelihood given a model that does not include a distribution for the missing data mechanism; examples of IL methods include ignorable maximum likelihood (IML), Bayesian inferences, or multiple imputation based on the predictive distribution from a Bayesian model, as in SAS PROC MI (SAS 2010) or IVEware (Raghunathan et al. 2001);

(c) Nonignorable modeling, which derives inference from the likelihood function based on a joint distribution of the variables and the missing data indicators (Little and Rubin 2002, chapter 15).

CC analysis is the default method in most software packages. Much of the statistical literature views CC with disfavor since it discards the incomplete cases.

However, CC has the advantage of yielding valid inference when the missingness of covariates does not depend on the outcome. This advantage of CC in regression analysis is usually overlooked.

Ignorable likelihood methods have the advantage of retaining all the data, but assume that missing data are missing at random (MAR), in the sense that missingness does not depend on missing values (Rubin 1976, Little and Rubin 2002). IL methods are fully efficient for well-specified models and they are also easy to fit since software packages are widely available (IVEWARE, PROC MI in SAS). Simulation studies show that IL methods are quite robust in the sense that it performs reasonably well even when the MAR assumption is slightly violated (Little and Zhang, 2011). This is because the efficiency gain by using more cases outweighs the bias resulting from incorrectly ignoring the missing data mechanism.

When the missingness of W is thought to depend on the missing value (MNAR), IL methods yield biased estimation. Nonignorable modeling methods, which jointly model the distribution of Y, W and R_w, were proposed (Lipsitz et al. 1999, Huang et al. 2005). There are several disadvantages with nonignorable modeling: (1) the model is not easy to specify correctly and sensitive to model misspecification; (2) the parameters might be inestimable and therefore the model usually needs restrictions to be identifiable; (3) there are limited software programs available for nonignorable modeling.

Other methods for handling missing covariates in regression include the indicator method and the stratification methods (Jones, 1996). In the missing-indicator method, an indicator of whether the covariate is missing is included in the regression model. The stratification methods divide the dataset into different strata for analysis. Both methods

avoid discarding the incomplete cases but might result in biased estimation of the regression coefficient and residual variance.

In Chapter 2, we review complete-case analysis (CC) and ignorable likelihood method (IL), and propose a hybrid class, subsample ignorable likelihood (SSIL) methods, which applies an IL method to the subsample of observations that are complete on one set of variables, but possibly incomplete on other variables and the outcome. Conditions on the missing data mechanism are presented under which SSIL gives consistent estimates, but both complete-case analysis and IL methods are inconsistent. We illustrate properties of the methods by simulation, and apply the proposed method to data from National Health and Nutrition Examination Survey and a liver cancer study. Extensions to non-likelihood analyses are also possible.

In Chapter 3, we consider the regression of outcome Y on regressors W and Z with some values of W missing, when our main interest is the effect of Z on Y, controlling for W. Besides the CC, IL, and NIM methods we discussed above, another simple practical approach that has not received much theoretical attention is to drop the regressor variables containing missing values from the regression modeling (DV, for drop variables). DV does not lead to bias when either (a) the regression coefficient of W is zero or (b) W and Z are uncorrelated. We propose a pseudo-Bayesian approach for regression with missing covariates that compromises between the CC and DV estimates, exploiting information in the incomplete cases when the data supports DV assumptions. We illustrate favorable properties of the method by simulation, and apply the proposed method to a liver cancer study. Extensions of the method to more than one missing covariates and to generalized linear models are also discussed.

In Chapter 4, we study the effect of covariate missingness on the estimation of the regression and answer the question when it is necessary to model the missing data mechanism. We will study two aspects of covariate missingness on the estimation of regression: (1) nonignorability, which concerns mainly how IL methods perform under varying levels of nonignorability; (2) outcome dependency, which studies the relatedness of covariate missingness to the outcome on the estimation of regression. We compare different methods for regression with missing covariates using a series of simulation experiments. We conclude the dissertation with a short discussion and future work in Chapter 5.

CHAPTER 2

Subsample Ignorable Likelihood for Regression with Missing Data

2.1 Introduction

Missing data is an important practical problem in many applications of statistics. We consider multivariate regression with missing data. Reviews of previous research on the topic include Little (1993), Ibrahim et al. (1999), Ibrahim et al. (2002), Ibrahim et al. (2005), and Chen et al. (2008). Three approaches are:

(a) Complete-case analysis (CC), which discards the incomplete cases;

(b) Ignorable likelihood (IL) methods, which base inferences on the observed likelihood given a model that does not include a distribution for the missing data mechanism; examples of IL methods include ignorable maximum likelihood (IML), Bayesian inferences, or multiple imputation based on the predictive distribution from a Bayesian model, as in SAS PROC MI (SAS 2010) or IVEware (Raghunathan et al. 2001);

(c) Nonignorable modeling, which derives inference from the likelihood function based on a joint distribution of the variables and the missing data indicators. This approach is less common in practice, because of the difficulty in specifying the model for missing data mechanism, sensitivity to misspecification of this distribution, problems with

identifying the parameters (Little and Rubin 2002, chapter 15), and lack of widely-available software.

IL methods have the advantage of retaining all the data, but assume the missing data are missing at random (MAR), in the sense that missingness of variables that contain missing values does not depend on the missing values, after conditioning on available data (Rubin 1976, Little and Rubin 2002). CC involves a loss of information, but has the advantage of yielding valid inferences when missingness depends on the missing covariates *X*'s but not the response *Y*, a potentially nonignorable mechanism where IL methods are subject to bias. This advantage of CC is sometimes overlooked in comparisons of the methods.

Can the information loss in CC analysis be mitigated, while retaining the useful property of allowing missingness to depend on the values of missing covariates? This article shows that the answer is yes, under particular assumptions about the missing data mechanism formalized in Section 2.4. The key idea is to divide the covariates into three sets – one set (say *Z*) fully observed, one set (say *W*) for which missingness is assumed to depend on *W* and other covariates but not on the outcomes *Y*, and a third set (say *X*), which together with *Y* are assumed MAR in the subsample of cases with *W* fully observed. The proposed method, subsample ignorable likelihood (SSIL), then applies an IL method to the subsample of cases with *W* observed. Particular forms discussed below are subsample ignorable maximum likelihood (SSIML), which applies IML to the subsample, and SSIMI, which applies an ignorable model to multiply-impute the missing values in the subsample.

Section 2.2 presents a motivating application based on data from the National Health and Nutrition Examination Survey (CDC 2004), where the regression of interest concerned the effect of income and education on blood pressure, adjusting for age, gender and body mass index (BMI). In this application, age and gender were fully observed, but the other variables had missing values; it was thought reasonable to assume missingness of education, BMI and the blood pressure measures was MAR, but missingness of income was thought likely to be dependent on income. Thus in this example, Z consists of age and gender, W consists of income, and X consists of education and BMI. The method consists of applying an IL method to the subset of cases with income observed. We formulate the problem in a way that encompasses multivariate regression and repeated measures analyses with missing data in outcomes and covariates.

Section 2.3 reviews properties of CC and IL, and Section 2.4 presents properties of the proposed SSIL methods. In particular, conditions on the missing data mechanism are presented under which SSIL gives consistent estimates, but both IL and CC analyses are inconsistent. In other circumstances, IL is inconsistent and SSIL and CC are consistent, but SSIL is more efficient than CC since it uses more of the data. Section 2.5 presents simulations that illustrate the properties of SSIL and alternative methods. In section 2.6 we apply the method to the motivating data from the National Health and Nutrition Examination Survey (NHANES) (CDC 2004). We conclude with some discussion in Section 2.7.

2.2 The motivating problem

The effect of socioeconomic status on blood pressure has been studied by many researchers (Gulliford et al. 2004, Colhoun et al. 1998, and etc). The results provide an important basis for public health interventions. The effect of socioeconomic status on blood pressure generally varies by geographical region and time as the risk factors in populations change (Mackenbach 1994). The data set analyzed in this article is from the 2003-2004 National Health and Nutrition Examination Survey (CDC 2004), a survey designed to assess the health and nutritional status of US adults and children. To study the effect of income and education on blood pressure, we extract the following data:

(a) two outcome measures: systolic blood pressure (SBP) and diastolic blood pressure (DBP);

(b) two socioeconomic status measures: household income (HHINC) and years of education (EDU, in years);

(c) three other covariates: age (in years), gender, and body mass index (BMI, kg/m^2).

Regressions of SBP and DBP on the covariates are fitted to study the effect of socioeconomic status on blood pressure.

Some of the variables have missing values -- see Table 2.1 for the proportion of missing values for each variable. CC analysis suffers from the loss of a large proportion of the cases. IL methods capture the partial information in the incomplete cases lost by CC analysis, but assume the missing values are MAR. It is reasonable to assume MAR for education, BMI, and the two blood pressure measures, but missingness of household income is thought more likely to be missing not at random (MNAR), since the probability of responding to income is thought likely to depend on the underlying value of income – often individuals with high or low values of income are considered less likely to respond

to income than others (David et al, 1986, Lillard et al. 1986, Yan et al. 2010). If these assumptions are correct, IL methods yield biased regression estimates. This motivates a new method which we call subsample ignorable likelihood (SSIL), which allows MAR assumptions for some variables (SBP, DBP, Education, BMI) and MNAR assumptions for others (Income), in a sense defined precisely in Section 2.4.

Before considering SSIL, it is useful to review more precisely the assumptions underlying IL and CC methods. This is the topic of the next section.

2.3 Complete-Case and Ignorable Likelihood Methods

In this section, we consider the data with the structure in Figure 2.1. Let $\{(z_i, w_i, y_i), i = 1,..n\}$ denote *n* independent observations on a (possibly multivariate) outcome variable *Y* and two sets of covariates, *Z* and *W*, where *Z* is fully observed and *W*, *Y* have missing values. Interest concerns the parameters ϕ of the distribution of *Y* given (*Z*, *W*), say $p(y_i \mid z_i, w_i, \phi)$.

The rows of Figure 2.1 divide the cases into two patterns. Pattern 1 ($i = 1,\ldots,m$) consists of complete cases, for which (z_i, w_i, y_i) are fully observed. Pattern 2 consists of cases where at least one of the variables in w_i, and possibly components of y_i, are missing. The column $R_{(w_i, y_i)}$ represents a vector of response indicators for (w_i, y_i), with entries 1 if a variable is observed and 0 if a variable is missing; R_{w_i} and R_{y_i} denotes the

Figure 2.1: General Missing Data Structure for Section 2.2

Pattern	Observation, i	z_i	w_i	y_i	$R_{(w_i,y_i)}$
1	$i = 1,\ldots,m$	√	√	√	$u_{(w,y)} = (1,\ldots,1)$
2	$i = m+1,\ldots,n$	√	x	?	$\bar{u}_{(w,y)}$

Key: √ denotes observed, x denotes at least one entry missing, ? denotes observed or missing

response indicators for w_i and y_i respectively. To describe missing data patterns for a set of variables (say v), it is convenient to write $u_v = (1,\ldots,1)$ to denote a vector of 1's of the same length as the vector v, and $\bar{u}_v$ to denote a vector of 0's and 1's of the same length as v for which at least one entry is zero. Then, for the cases i in Figure 2.1, $R_{(w_i,y_i)} = u_{(w,y)}$ for the complete cases in Pattern 1 and $R_{(w_i,y_i)} = \bar{u}_{(w,y)}$ for the incomplete cases in Pattern 2. The pattern of missing values will typically vary over these cases, but we do not need to distinguish them for the present discussion.

IL inference requires a model for the distribution of W and Y given Z indexed by parameters θ, say $p(w_i, y_i \mid z_i, \theta)$ -- the fully observed covariates can be treated as fixed

(Little & Rubin 2002, Section 11.4.) The ignorable likelihood is obtained by integrating the missing variables out of this joint distribution, and treating θ as the argument of the resulting density. That is:

$$L_{\text{ign}}(\theta) = \text{const.} \times \prod_{i=1}^{n} p(w_{\text{obs},i}, y_{\text{obs},i} \mid z_i, \theta), \tag{2.1}$$

where $(w_{\text{obs},i}, y_{\text{obs},i})$ are the observed components of (w_i, y_i), respectively. For Bayesian inferences this likelihood is multiplied by a prior distribution for θ. Inferences about the parameter $\phi = \phi(\theta)$ of interest are obtained from inferences of θ in the usual way. In particular, the ML estimate is $\hat{\phi} = \phi(\hat{\theta})$ where $\hat{\theta}$ is the ML estimate of θ, and draws from the posterior distribution of ϕ are $\phi^{(d)} = \phi(\theta^{(d)})$, where $\theta^{(d)}$ is a draw from the posterior distribution of θ. Rubin's (1976) theory shows that a sufficient condition for valid inferences based on (1) is that the data are missing at random (MAR), that is:

$$p(R_{w_i}, R_{y_i} \mid z_i, w_i, y_i, \psi) = p(R_{w_i}, R_{y_i} \mid z_i, w_{\text{obs},i}, y_{\text{obs},i}, \psi) \ , \tag{2.2}$$

where ψ are parameters for the missing data mechanism. If, in addition, the parameters θ and ψ are distinct, inferences based on (2.1) are fully efficient; but MAR is the important condition in practice.

CC analysis bases inferences for ϕ on the complete observations in Pattern 1. In a likelihood context, the method bases inference on the conditional likelihood corresponding to the complete cases, namely:

$$L_{\text{cc}}(\phi) = \text{const.} \times \prod_{i=1}^{m} p \ \ y_i \mid w_i, z_i, R_{(w_i, y_i)} = u_{(w,y)}; \phi \ , \tag{2.3}$$

The key condition under which inference based on $L_{cc}(\phi)$ is valid is that the probability that an observation is complete does not depend on the outcomes, that is:

$$p\ \ R_{(w_i,y_i)} = u_{(w,y)} \mid z_i, w_i, y_i, \psi) \ = p\ \ R_{(w_i,y_i)} = u_{(w,y)} \mid z_i, w_i, \psi) \quad \text{for all } y_i \qquad (2.4)$$

Figure 2.2: Missing Data Pattern of Example 2.1

Pattern	Observation, i	z_i	w_i	y_i	$R_{(w_i,y_i)}$
1	$i = 1,\ldots,m$	√	√	√	(1,1)
2	$i = m+1,\ldots,n$	√	x	√	(0,1)

Key: √ denotes observed, x denotes missing

Note that this condition allows missingness to be MNAR, since missingness can depend on the values of W which are sometimes missing. CC analysis works in this case because Eq. (2.4) implies that

$$p\ \ y_i \mid w_i, z_i, R_{(w_i,y_i)} = u_{(w,y)}, \phi \ = p\ \ y_i \mid w_i, z_i, \phi \ \ ,$$

so the regression based on the complete cases is the regression of interest for the whole sample. The likelihood for a fully specified model with parameters (ϕ, ψ) can be written as

$$L(\phi, \gamma \mid Z, W_{\text{obs}}, Y_{\text{obs}}, R_{(w,y)}) = L_{cc}(\phi) L_{\text{rest}}(\phi, \psi),$$

and the component $L_{\text{rest}}(\phi,\psi)$ is discarded. ML estimates based on $L_{\text{cc}}(\phi)$ are consistent and asymptotically normal, but are not necessarily fully efficient, since $L_{\text{rest}}(\phi,\psi)$ may contain information about the parameters of interest ϕ. However, recovering this information requires a model for the missing data mechanism, which may be difficult to specify correctly, and which is not needed for CC analysis.

Example 2.1. Missing data in a single covariate. Figure 2.2 displays a special case of Figure 2.1 where w_i and y_i are single variables, and the incomplete cases have w_i missing (denoted x) but not y_i. The MAR condition (2.2) becomes

$$p(R_{(w_i,y_i)}=(1,1)\,|\,z_i,w_i,y_i,\psi)=p(R_{(w_i,y_i)}=(1,1)\,|\,z_i,y_i,\psi) \text{ for all } w_i, \qquad (2.5)$$

and (2.4) becomes

$$p(R_{(w_i,y_i)}=(1,1)\,|\,z_i,w_i,y_i,\psi)=p(R_{(w_i,y_i)}=(1,1)\,|\,z_i,w_i,\psi) \text{ for all } y_i. \qquad (2.6)$$

The choice between IL or CC rests on whether (2.5) or (2.6) is a better assumption for the missing data mechanism, that is, on whether missingness of *W* is thought to depend on *Y* and *Z* (but not *W*) or on *W* and *Z* (but not *Y*). Little and Wang (1996, Example 2) presents a normal pattern-mixture model where missingness is a function of $w_i+\lambda y_i$, for which the ML estimates correspond to IL when $\lambda=0$ and CC when $\lambda=\infty$. An interesting feature of that example is that CC analysis is not just consistent but also fully efficient under (2.6).

We note that CC analysis is viewed with disfavor in the missing data literature, because of the loss of information in the incomplete cases. Many simulation studies in the literature (e.g. Little 1979, Chen, Zeng and Ibrahim 2007) show superiority of IL over CC,

but are biased towards IL because they are based on MAR data. The above arguments also apply to repeated measures models where *Y* is multivariate and both *Y* and covariates contain missing values. In this setting, CC is still a superior alternative to IL if missingness depends on covariates, including those with missing values, but not on the repeated measures *Y*. We are not aware of this advantage of CC being considered in the repeated-measures setting, where attention has been focused on capturing the information

Figure 2.3: General Missing Data Structure for Section 2.3

Pattern	Observation, i	z_i	w_i	x_i	y_i	R_{w_i}	$R_{(x_i,y_i)}$
1	$i = 1,\ldots,m$	√	√	√	√	u_w	$u_{(x,y)}$
2	$i = m+1,\ldots,m+r$	√	√	?	?	u_w	$\bar{u}_{(x,y)}$
3	$i = m+r+1,\ldots,n$	√	x	?	?	$\bar{u}_w$	$u_{(x,y)}$ or $\bar{u}_{(x,y)}$

Key: √ denotes observed, x denotes at least one entry missing, ? denotes observed or missing

in the incomplete cases.

2.4 Subsample Ignorable Likelihood Methods -- Theory

We consider the missing data pattern in Figure 2.3, in which another set of incomplete covariates X is added. The observations are grouped into three patterns: Pattern 1 consists of the complete cases ($R_{w_i} = u_w$, $R_{(x_i,y_i)} = u_{(x,y)}$), Pattern 2 incomplete cases with W fully observed ($R_{w_i} = u_w$, $R_{(x_i,y_i)} = \bar{u}_{(x,y)}$), and Pattern 3 cases with W incomplete ($R_{w_i} = \bar{u}_w$). Interest concerns the parameters ϕ of the distribution of Y given (Z, W, X), say $p(y_i \mid z_i, w_i, x_i, \phi)$. We propose subsample IL (SSIL), which applies an IL method to the subsample of cases in Patterns 1 and 2 with both Z and W observed.

The division of covariates into W and X for SSIL is determined by assumptions about the missing data mechanism. Specifically, the method is valid under the following two assumptions:

(a) Covariate missingness of *W*: the probability that *W* is fully observed depends only on the covariates and not *Y*, that is:

$$p\ \ R_{w_i} = u_w \mid z_i, w_i, x_i, y_i, \psi_w) \ = p\ \ R_{w_i} = u_w \mid z_i, w_i, x_i, \psi_w) \quad \text{for all } y_i \tag{2.7}$$

(b) Subsample MAR of *X, Y*: Missingness of *X* and *Y* is MAR within the subsample of cases for which *W* is fully observed, that is:

$$\begin{aligned} &p(R_{(x_i, y_i)} \mid z_i, w_i, x_i, y_i, R_{w_i} = u_w; \psi_{xy \cdot w}) = \\ &\qquad p(R_{(x_i, y_i)} \mid z_i, w_i, x_{\text{obs},i}, y_{\text{obs},i}, R_{w_i} = u_w; \psi_{xy \cdot w}) \quad \text{for all } x_{\text{mis},i}, y_{\text{mis},i} \end{aligned} \tag{2.8}$$

To establish the validity of SSIL under (2.7) and (2.8), we first consider the conditional likelihood for a set of parameters ζ based on the joint distribution of $X, Y, R_{(X,Y)}$ given *W* and *Z* and $R_{w_i} = u_w$, that is, restricted to cases *i* with *W* fully observed:

$$L_{\text{cc,w}}(\zeta) = \prod_{i=1}^{m+r} p\ \ x_{\text{obs},i}, y_{\text{obs},i}, R_{(x_i, y_i)} \mid w_i, z_i, R_{w_i} = u_w; \zeta \ ,$$

where $\zeta = (\theta, \psi)$. By a direct application of Rubin's (1976) theory, under the subsample MAR condition (2.8), this likelihood factorizes as

$$L_{\text{cc,w}}(\zeta) = \prod_{i=1}^{m+r} p\ \ x_{\text{obs},i}, y_{\text{obs},i} \mid w_i, z_i, R_{w_i} = u_w; \theta \ \times \prod_{i=1}^{m+r} p\ \ R_{(x_i, y_i)} \mid w_i, x_{\text{obs},i}, y_{\text{obs},i}, z_i, R_{w_i} = u_w; \psi \ ,$$

where the second component on the right side does not involve θ, and the first component on the right side, namely

$$L_{\text{ign,w}}(\theta) = \prod_{i=1}^{m+r} p\ \ x_{\text{obs},i}, y_{\text{obs},i} \mid w_i, z_i, R_{w_i} = u_w; \theta \ ,$$

is the likelihood for the subsample with w_i observed, ignoring the distribution of the missing data indicators $R_{(x_i,y_i)}$. Thus inference about θ, the parameter of the distribution (*X*,*Y*) given (*W*, *Z*), based on $L_{\text{ign,w}}(\theta)$ is valid. Now factorize

$$p\ x_i, y_i \mid w_i, z_i, R_{w_i} = u_w; \theta\ =$$
$$p\ y_i \mid x_i, w_i, z_i, R_{w_i} = u_w; \theta\ \times p\ x_i \mid w_i, z_i, R_{w_i} = u_w; \theta\ .$$

By assumption (2.7), $p\ y_i \mid x_i, w_i, z_i, R_{w_i} = u_w; \theta\ = p\ y_i \mid x_i, w_i, z_i, \phi$, where $\phi = \phi(\theta)$ is the parameter of the regression of interest, and the conditioning on the cases with *W* observed is removed. Thus, under assumptions (2.7) and (2.8), we can base inferences about θ on $L_{\text{ign,w}}(\theta)$, and then derive likelihood inferences about $\phi = \phi(\theta)$ as in Section 2.2.

The missing data mechanism defined by conditions (2.7) and (2.8) is suitable in empirical studies where it is natural to assume covariate-dependent missingness for some covariates and subsample MAR missingness for others. For example, in the motivating example concerning the regression of blood pressure on socioeconomic variables in Section 2.2, Income may be covariate-dependent and the Education and BMI may be subsample MAR. In environmental health research, values of variables that are missing because they lie below the limit of detection (LOD) are MNAR. If missing values exist for other variables and can be assumed to be MAR, then SSIL on the subsample with measurements within the detection limit yields valid regression inference.

Generally, SSIL methods are based on a partial likelihood (Cox 1972) with the component $L_{\text{ign,w}}(\theta)$ discarded from the analysis and hence involve a loss of efficiency

Figure 2.4: Missing Data Structure for Example 2.2

Pattern	Observation, i	z_i	w_i	x_i	y_i	R_{w_i}	R_{x_i}
1	$i = 1,\ldots,m$	√	√	√	√	1	1
2	$i = m+1,\ldots,m+r$	√	√	x	√	1	0
3	$I = m+r+1,\ldots,n$	√	x	√	√	0	1

Key: √ denotes observed, x denotes missing.

relative to full likelihood methods. However, they are more efficient than CC analysis, and avoid the need to specify the form of the missing data mechanism beyond assumptions (2.7) and (2.8).

Assumptions (2.7) and (2.8) differ from the assumptions under which IL and CC methods are valid. Specifically, IL inference assumes the data are MAR, that is:

$$p(R_{w_i}, R_{(x_i,y_i)} \mid z_i, w_i, x_i, y_i, \psi) = p(R_{w_i}, R_{(x_i,y_i)} \mid z_i, w_{\text{obs},i}, x_{\text{obs},i}, y_{\text{obs},i}, \psi) \quad \text{for all } w_{\text{mis},i}, x_{\text{mis},i}, y_{\text{mis},i} \tag{2.9}$$

This differs from conditions (2.7) and (2.8), where missingness of both w_i and (x_i, y_i) can depend on missing components of w_i. CC analysis yields valid inferences if the probability that an observation is complete does not depend on the outcomes, that is:

$$p\left(R_{w_i} = u_w, R_{(x_i,y_i)} = u_{(x,y)} \mid z_i, w_i, x_i, y_i, \psi\right) = p\left(R_{w_i} = u_w, R_{(x_i,y_i)} = u_{(x,y)} \mid z_i, w_i, x_i, \psi\right) \quad \text{for all } y_i \tag{2.10}$$

This differs from the assumption (2.8) in that missingness of (x_i, y_i) in (2.8) can depend on the observed components of y_i. If this is not the case, then CC yields valid inferences but is less efficient than SSIL, since SSIL uses the data in Pattern 2, which is discarded by CC.

Example 2.2: a normal regression model with two incompletely observed covariates

Figure 4 displays a special case of Figure 3, where *W*, *X* and *Y* (but not necessarily *Z*) are univariate, Z and Y are fully observed, *X* is missing and *W* is observed in Pattern 2, and *W* is missing and *X* is observed in Pattern 3. Restating assumptions (2.7) and (2.8) in this special case yields:

$$p\ R_{w_i} = 1 | z_i, w_i, x_i, y_i, \psi_w\ = p\ R_{w_i} = 1 | z_i, w_i, x_i, \psi_w \quad \text{for all } y_i \tag{2.11}$$

$$p(R_{x_i} = 1 | z_i, w_i, x_i, y_i, R_{w_i} = 1, \psi_{xy \cdot w}) = p(R_{x_i} = 1 | z_i, w_i, y_i, R_{w_i} = 1, \psi_{xy \cdot w}) \text{ for all } x_i \tag{2.12}$$

Under this mechanism, SSIL yields consistent estimates, but (a) CC analysis may yield inconsistent estimates since missingness of *X* may depend on the outcome *Y*, and (b) IL methods may yield inconsistent estimates, since missingness of *W* can depend on missing values of *W* (i.e. MNAR).

2.5 Simulation Study

As a numerical illustration of the theory in Section 2.4, we simulate data for the pattern of Example 2.2, under a variety of missing data mechanisms. For each of 1000 replications, 1000 observations (z_i, w_i, x_i, y_i) , i = 1,…,1000 on *Z*, *W*, *X* and *Y* were generated as follows:

$$(z_i, w_i, x_i) \sim_{\text{ind}} N(0, \Sigma),$$

where $N(\mu, \Sigma)$ denotes the normal distribution with mean μ and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

and

$$y_i \mid z_i, w_i, x_i \sim_{\text{ind}} N(1 + z_i + w_i + x_i, 1) .$$

Missing values of *W* and *X* were then generated from the following two logistic models:

$$\text{logit } P(R_{w_i} = 0 \mid z_i, w_i, x_i, y_i) = \alpha_0^{(w)} + \alpha_z^{(w)} z_i + \alpha_w^{(w)} w_i + \alpha_x^{(w)} x_i + \alpha_y^{(w)} y_i$$
$$\text{logit } P(R_{x_i} = 0 \mid R_{w_i} = 1, z_i, w_i, x_i, y_i) = \alpha_0^{(x)} + \alpha_z^{(x)} z_i + \alpha_w^{(x)} w_i + \alpha_x^{(x)} x_i + \alpha_y^{(x)} y_i$$

with x_i fully observed when w_i is missing.

For the missing data generation schemes above, CC analysis is valid if both $\alpha_y^{(w)}$ and $\alpha_y^{(x)}$ are zero; IL is valid if $\alpha_w^{(w)}$, $\alpha_x^{(w)}$ and $\alpha_x^{(x)}$ are zero; SSIL is valid if $\alpha_y^{(w)}$ and $\alpha_x^{(x)}$ are zero. Four missing data mechanisms were created using different sets of values for the regression coefficients such that, in mechanism (I) all three methods (CC, IL and SSIL) are consistent, while in mechanisms (II), (III) and (IV), just one of the three methods is valid. The simulation setup is summarized in Table 2.2.

These missing data mechanisms all generate from 20% to 35% of values missing in *W* and *X*, respectively. Three values of the correlation of *X* and *W*, $\rho = 0,\ 0.3$ and 0.8, are chosen, to examine the impact of correlation between the covariates.

Four specific versions of the methods are applied to estimate the regression coefficients:

(1) CC: Complete-case analysis, using ordinary least squares;

(2) IML: ignorable ML for the whole dataset;

(3) SSIML: IML for the subsample with W observed;

(4) BD: least squares estimates from the regression before deletion (BD), as a benchmark method.

For each method, Table 2.3 summarizes the root mean squared errors (RMSEs) of estimates of all the regression coefficients, and Tables 2.4, 2.5 and 2.6 report respectively the empirical bias, RMSE and coverage probability of estimates of the individual regression coefficients. Results in bold type reflect situations where the method is consistent based on the theory of Section 2.4, and hence should do well. The results are based on 1000 repetitions in each simulation.

In general, the simulation results are in line with theoretical expectations. Results for SSIML lie between those for CC and IML for mechanisms I, II and III, where one or both of CC and IML are consistent – both CC and IML in mechanism I, CC in mechanism II and IML in mechanism III. This finding reflects the fact that SSIML is a hybrid of CC and IML, sharing features of both methods. In mechanism IV, SSIML is consistent but CC and IML are inconsistent, and in this case SSIML has small empirical bias and generally performs best, except for some individual coefficients where the gain in efficiency of IML compensates for the bias of that method. We now describe results in a bit more detail.

For mechanism I, all three methods yield consistent estimates, IML is best since it makes full use of the data, CC is the worst since it discards the most information, and

SSIML lies between CC and IML, since it retains some incomplete cases and drops others.

For mechanism II, CC is valid and in general has the lowest RMSEs, while both IML and SSIML are biased, with SSIML having RMSEs lying between those of CC and IML. However, for $\rho = 0.8$, SSIML and IML yield comparable or even smaller RMSEs than CC for β_z and β_w, reflecting gains in efficiency that compensate for bias in these parameter estimates.

For mechanism III, IML is the only valid method among the three, and is clearly the best method. Both CC and SSIML lead to biased estimates, as shown in Table 2.3, with SSIML being better than CC since it is incorporates features of IML as a method.

In mechanism IV, SSIML is valid while CC and IML are biased. The RMSEs from SSIML are generally the smallest, except that IML yields a smaller RMSE than SSIML for β_w.

In some of these situations, supporters of IML may note that it competes well with other methods, despite its theoretical inconsistency and the quite sizeable sample size. This suggests a degree of robustness for IML, which has the virtue of retaining all the data.

2.6 Application to motivating example

We now apply the proposed method to the NHANES (2003-2004) data presented in Section 2.2. Two blood pressure measurements: systolic blood pressure (SBP) and diastolic blood pressure (DBP), are regressed on household income (HHINC, in dollars/yr) and years of education (EDU, in years), adjusting for age (in years), gender

and body mass index (BMI, kg/m^2). Household income data are categorical with 11 categories in the NHANES, and we use the median of the corresponding category as a proxy to the true household income. Education is dichotomized to be high-school and above vs. less than high-school.

Age and gender are fully observed, while household income, education, BMI and the two blood pressure measures are subject to missing data, with the percentages shown in Table 1. We assume covariate missingness for household income, given evidence that people with high or low income are more likely to fail to report it, and assume subsample MAR for other variables: (1) missingness of BMI and blood pressure measurements is likely missing completely at random due to missing visit; (2) with income observed, it is reasonable to assume MAR for education because income and education are correlated (Tolley and Olson, 1971). With these two plausible assumptions, SSIL on the subsample with household income observed yields consistent estimates of the regression, while IL on the whole sample may be biased. CC analysis is also valid since there is little evidence to believe that missingness of covariates depends on blood pressure; however, SSIL is preferred over CC since it uses more information in the incomplete cases than CC analysis. For simplicity, we ignore the design features (weighting and clustering, etc) of the NHANES study. For the SSIL method, we use IVEware to multiply impute missing values in the subsample with household income observed, and then use SAS software (SAS 2010) to perform the regression analyses and to combine results from individual imputed dataset. We denote this method SSIMI. For the IL method, we use IVEware to multiply impute the full sample, and use SAS software for regression analyses and combining the results. We denote this method IMI. The results of CC analysis, SSIMI

analysis and IMI are shown in Table 2.6. All three methods yield similar estimates of the effect of household income on blood pressure, statistically not significant for SBP but significant for DBP, with blood pressure increasing with income. There is a negative association between education and SBP and a positive association between education and DBP, regardless of method of analysis. For education, SSIMI and CC yield similar and stronger effects on the two blood pressure measures than IMI, implying possible bias in IMI given the above assumptions about the missing data mechanism. The larger sample of SSIMI over CC should result in a gain in efficiency for SSIMI in this situation, although CC and SSIMI have similar estimated standard errors for this particular sample.

2.7 Discussion

The idea behind SSIL, to apply an analysis that assumes MAR to a subsample of the data that is complete on a subset of the covariates, is both simple and powerful. SSIL analysis has the following strengths: (1) It is easy to implement, since existing software for doing MAR analyses is all that is required, and this software is now widely available for many common models; (2) It avoids discarding all incomplete cases, thus alleviating one of the drawbacks of CC analysis; (3) It applies to a broad class of univariate and multivariate regression models, including multivariate linear regression, generalized linear models (GLMs) and generalized linear mixed models (GLMMs); and (4) The method works for a class of missing data mechanisms, defined by (2.7) and (2.8), where both IL and CC methods fail to give consistent estimates. This extends the class of MNAR models that can be handled by a selective use of MAR methods, and allows combinations of MAR and MNAR mechanisms for different variables in the data set.

In another analysis which drops a subset of incomplete cases, Von Hippel (2007) applies an MAR multiple imputation analysis in the regression setting, where a univariate outcome Y has missing values, and then applies the final regression analysis to the subsample of cases with Y observed, that is, dropping the cases with Y imputed. This strategy reduces the simulation error from multiple imputation, but it is applied within a univariate regression for a MAR model, and hence is much less general than SSIL, and does not generate a method that is consistent for a MNAR mechanism.

The general theoretical rationale of SSIL is partial likelihood (Cox, 1972). This involves a potential loss of efficiency relative to full modeling, but it is much simpler, since the latter requires specifying the precise form of the missing data mechanism via a model for the missing data indicators, which is vulnerable to model misspecification. Also, existing software for full MNAR models is not widely available.

An important topic is how much efficiency is lost by SSIL relative to full likelihood methods. SSIL involves minimal loss when the fraction of cases in the subsample with the MNAR subset W observed is relatively high, and hence the method is most beneficial relative to CC when the fraction of information in the pattern with W complete but other variables incomplete is relatively high. It can be shown by an extension of the arguments in Little and Wang (1996) that for the data in Example 2, the SSIL method is in fact full ML for a particular normal pattern-set mixture model (Little 1993). This aspect of SSIL methods will be the subject of a future paper.

The form of IL method in SSIL is left unspecified in this article where possible, for increased generality. As noted, options for IL include maximum likelihood (IML), multiple imputation using software like PROC MI or IVEware (Raghunathan et al. 2001),

and fully Bayes methods using software such as BUGS (Gilks et al. 1994). Mixing these methods is also advantageous in some settings.

The idea of SSIL is presented here in the context of likelihood-based analyses, but it also applies to non-likelihood analyses that are valid under the MAR assumption. For example, for repeated-measures data, the IL method applied to the subsample could be replaced by a method such as weighted generalized estimating equations (WGEE), which is also valid under MAR, without affecting the validity of the method under the stated assumptions (2.7) and (2.8).

From a practitioner's viewpoint, the main challenge in applying SSIL is deciding which covariates belong in the set W and which belong in the set X; that is, which covariates are used to create the subsample for the MAR analysis. The choice is guided by the basic assumptions (2.7) and (2.8), concerning which variables are considered covariate-dependent MNAR and which are considered subsample MAR. This is a substantive choice that requires an understanding about the missing data mechanism in the particular context. It is aided by learning more about the missing data mechanism, for example by recording reasons why particular values are missing. Although a challenge, we note that the same challenge is present in any missing data method, including CC, IL and WGEE. When faced with missing data, assumptions are inevitable, and they need to be as reasonable and well-considered as possible.

In cases where a choice cannot be made, an alternative strategy is simply to see whether key results are robust to alternative methods. Thus, one might apply CC, IL and SSIL for subsamples judiciously chosen based on assumptions (2.7) and (2.8), to assess

sensitivity of key inferences to alternative assumptions about the missing-data mechanism.

Table 2.1: Percentages of Missing Data in NHANES[a] 2003-2004

Partition[b]	Variables	Full Data (n=9041)	Subset with HHINC[c] observed (n=5400)
W	HHINC[c] (1k dollars/ yr)	40.27	0
	Age (years)	0	0
Z	Gender	0	0
	Education (years)	17.24	16.74
X	BMI[c] (kg/m^2)	9.84	9.48
	SBP[c] (mmHg**)**	25.02	24.5
Y	DBP[c] (mmHg**)**	25.02	24.5

[a]:NHANES: National Health and Nutrition Examination Survey
[b]: Partition based on covariate missingness and subsample MAR
[c]: HHINC: household income; SBP: systolic blood pressure; DBP: diastolic blood pressure

Table 2.2: Missing data mechanisms generated in the simulations

Mechanisms	$\alpha_0^{(w)}$	$\alpha_z^{(w)}$	$\alpha_w^{(w)}$	$\alpha_x^{(w)}$	$\alpha_y^{(w)}$	$\alpha_0^{(x)}$	$\alpha_z^{(x)}$	$\alpha_w^{(x)}$	$\alpha_x^{(x)}$	$\alpha_y^{(x)}$
I: All valid	-1	1	0	0	0	-1	1	0	0	0
II: CC valid	-1	1	1	1	0	-1	1	1	1	0
III: IML valid	-2	1	0	0	1	-2	1	1	0	1
IV: SSIML valid	-1	1	1	1	0	-2	1	1	0	1

Missing value of W and X are generated based on the following logistic models:

$$\text{logit}\ \ P(R_{w_i} = 0 \mid z_i, w_i, x_i, y_i) = \alpha_0^{(w)} + \alpha_z^{(w)} z_i + \alpha_w^{(w)} w_i + \alpha_x^{(w)} x_i + \alpha_y^{(w)} y_i \quad .$$

$$\text{logit}\ \ P(R_{x_i} = 0 \mid R_{w_i} = 1, z_i, w_i, x_i, y_i) = \alpha_0^{(x)} + \alpha_z^{(x)} z_i + \alpha_w^{(x)} w_i + \alpha_x^{(x)} x_i + \alpha_y^{(x)} y_i$$

In particular, for the four missing data mechanisms:
I: Missingness of $W = f(Z)$, Missingness of $X = f(Z|W$ observed), all four methods are valid;
II: Missingness of $W = f(Z,W, X)$, Missingness of $X = f(Z,W,X|W$ observed), only CC valid;
III: Missingness of $W = f(Z)$, Missingness of $X = f(Z,W|W$ observed), only IML valid;
IV: Missingness of $W= f(Z,W,Y)$, Missingness of $X = f(Z,W,Y|W$ observed), only SSIML valid.

Table 2.3: Summary RMSEs*1000 of Estimated Regression Coefficients for Before Deletion (BD), Complete Cases (CC), Ignorable Maximum Likelihood (IML) and Subsample Ignorable Maximum Likelihood (SSIML), under Four Missing Data Mechanisms

	$\rho = 0$				$\rho = 0.3$				$\rho = 0.8$			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
BD	**65**	**64**	**63**	**64**	**67**	**68**	**67**	**67**	**106**	**106**	**107**	**106**
CC	**116**	**104**	555	335	**121**	**109**	503	296	**179**	**113**	450	285
IML	**83**	144	**83**	140	**84**	210	**84**	137	**133**	195	**128**	361
SSIML	**103**	159	368	**99**	**106**	144	356	**105**	**151**	130	346	**152**

*Four missing data mechanisms:
I: Missingness of $W = f(Z)$, Missingness of $X = f(Z|W$ observed), all four methods are valid;
II: Missingness of $W = f(Z,W, X)$, Missingness of $X = f(Z,W,X|W$ observed), only CC valid;
III: Missingness of $W = f(Z)$, Missingness of $X = f(Z,W|W$ observed), only IML valid;
IV: Missingness of $W= f(Z,W,Y)$, Missingness of $X = f(Z,W,Y|W$ observed), only SSIML valid.

RMSE estimates $1000 * \sqrt{E \left\| \beta_r - \beta_{TRUE} \right\|^2}$, with r denoting the r^{th} repetition.

Bold values are for methods consistent for the mechanism generating the data

Table 2.4: Empirical Bias*1000 for Individual Regression Coefficients under Four Missing Data Mechanisms (1000 replications)

$\rho = 0$																
	Mechanism I				Mechanism II				Mechanism III				Mechanism IV			
Method	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x
BD	**1**	**-1**	**0**	**1**	**-1**	**0**	**0**	**-1**	**-2**	**0**	**1**	**-1**	**0**	**-1**	**0**	**-2**
CC	**3**	**-1**	**-1**	**2**	**0**	**-1**	**2**	**-2**	-454	-229	-154	-115	-259	-123	-123	-61
IML	**3**	**-1**	**1**	**1**	204	63	41	73	**-2**	**0**	**3**	**-1**	99	31	7	44
SSIML	**4**	**0**	**-1**	**4**	112	37	39	19	-290	-170	-83	-85	**-19**	**-15**	**-14**	**-5**
$\rho = 0.3$																
Method	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x
BD	**0**	**-1**	**1**	**0**	**0**	**-1**	**2**	**0**	**-2**	**2**	**0**	**0**	**0**	**1**	**-1**	**3**
CC	**1**	**0**	**0**	**2**	**2**	**1**	**1**	**0**	-427	-182	-132	-91	-238	-95	-95	-42
IML	**0**	**0**	**2**	**2**	168	47	26	58	**-3**	**0**	**1**	**0**	93	33	3	46
SSIML	**1**	**-1**	**0**	**5**	97	30	31	14	-292	-145	-74	-74	**-25**	**-14**	**-16**	**1**
$\rho = 0.8$																
Method	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x
BD	**1**	**0**	**1**	**-1**	**-2**	**2**	**-2**	**0**	**0**	**2**	**-3**	**1**	**0**	**-4**	**2**	**3**
CC	**1**	**0**	**4**	**-4**	**0**	**4**	**-2**	**-1**	-382	-135	-100	-67	-212	-74	-68	-27
IML	**2**	**-5**	**1**	**-4**	89	35	20	44	**2**	**2**	**-3**	**1**	48	3	-2	40
SSIML	**0**	**-3**	**3**	**2**	41	16	19	8	-279	-117	-62	-56	**-19**	**-15**	**-10**	**8**

Table 2.5: RMSE*1000 for Individual Regression Coefficients under Four Missing Data Mechanisms (1000 replications)

$\rho = 0$																
	Mechanism I				Mechanism II				Mechanism III				Mechanism IV			
Method	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x
BD	**32**	**33**	**32**	**32**	**32**	**32**	**32**	**32**	**32**	**30**	**32**	**32**	**32**	**33**	**31**	**32**
CC	**57**	**57**	**54**	**63**	**57**	**49**	**50**	**52**	457	234	171	125	265	134	133	78
IML	**40**	**41**	**42**	**42**	209	75	58	84	**45**	**40**	**39**	**41**	108	51	40	60
SSIML	**50**	**52**	**50**	**53**	124	59	61	53	294	175	95	96	**54**	**50**	**48**	**45**
$\rho = 0.3$																
Method	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x
BD	**32**	**34**	**36**	**32**	**31**	**35**	**35**	**35**	**30**	**35**	**33**	**35**	**33**	**33**	**34**	**34**
CC	**56**	**62**	**65**	**59**	**61**	**52**	**51**	**53**	431	190	142	103	245	108	108	66
IML	**37**	**43**	**45**	**43**	173	62	72	72	**43**	**44**	**39**	**42**	102	53	42	62
SSIML	**48**	**53**	**58**	**52**	111	56	52	52	296	154	87	87	**60**	**51**	**50**	**48**
$\rho = 0.8$																
Method	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x
BD	**31**	**59**	**59**	**58**	**31**	**59**	**60**	**57**	**32**	**59**	**59**	**59**	**31**	**58**	**58**	**59**
CC	**53**	**99**	**99**	**99**	**62**	**53**	**55**	**56**	387	160	129	104	222	111	109	88
IML	**39**	**72**	**73**	**75**	158	54	74	67	**38**	**69**	**74**	**68**	118	285	135	130
SSIML	**47**	**81**	**85**	**83**	97	53	47	50	284	141	99	96	**56**	**80**	**81**	**83**

Table 2.6: 95% Confidence Coverage for Individual Regression Coefficients under Four Missing Data Mechanisms (1000 replications)

$\rho = 0$																
	Mechanism I				Mechanism II				Mechanism III				Mechanism IV			
Method	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x
BD	**95.2**	**94.3**	**95.1**	**95.0**	**95.4**	**95.0**	**95.5**	**95.3**	**93.9**	**96.3**	**94.6**	**94.9**	**94.5**	**93.3**	**95.0**	**95.5**
CC	**94.3**	**94.8**	**95.3**	**94.2**	**94.7**	**95.1**	**94.0**	**94.8**	0	0.6	7.9	28.0	0.9	32.5	32.3	75.3
IML	**94.3**	**94.2**	**94.0**	**94.5**	0.3	63.6	81.2	56.4	**93.7**	**95.5**	**94.2**	**94.0**	32.5	86.8	94.5	77.5
SSIML	**94.6**	**94.6**	**93.9**	**94.0**	38.5	86.7	85.3	91.1	0	4.5	52.5	48.3	**92.6**	**93.2**	**94.5**	**94.6**
$\rho = 0.3$																
Method	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x
BD	**95.0**	**95.0**	**94.1**	**95.0**	**95.7**	**95.1**	**95.2**	**94.1**	**95.7**	**93.7**	**95.6**	**94.8**	**94.5**	**95.5**	**94.7**	**95.8**
CC	**95.2**	**95.2**	**94.0**	**96.0**	**93.9**	**94.8**	**94.4**	**95.4**	0	8.1	25.6	52.3	2.9	56.7	57.5	85.4
IML	**94.1**	**94.1**	**94.0**	**95.8**	1.8	79.0	90.1	70.9	**94.9**	**94.2**	**95.6**	**93.2**	38.9	86.7	94.9	78.1
SSIML	**94.5**	**94.5**	**93.6**	**96.5**	53.3	90.1	89.7	93.7	0.1	20.0	65.9	64.0	**91.9**	**94.6**	**94.2**	**94.3**
$\rho = 0.8$																
Method	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x	β_0	β_z	β_w	β_x
BD	**95.4**	**95.3**	**94.7**	**94.7**	**95.1**	**94.3**	**94.0**	**95.5**	**95.0**	**95.3**	**94.3**	**95.1**	**96.2**	**95.3**	**95.2**	**94.6**
CC	**94.7**	**95.2**	**94.0**	**94.4**	**94.7**	**95.9**	**95.1**	**95.2**	0.0	61.7	77.7	87.7	10.2	86.5	87.0	94.3
IML	**96.2**	**95.9**	**94.3**	**95.4**	56.8	93.2	95.2	94.6	**94.3**	**93.6**	**94.5**	**94.7**	84.6	96.1	98.0	96.6
SSIML	**95.9**	**95.2**	**94.7**	**94.1**	89.1	94.9	95.5	94.0	0.1	68.7	86.9	90.4	**93.0**	**94.7**	**94.4**	**95.4**

Table 2.7: Estimates of the Effect of Socieconomic Status on Blood Pressure (NHANES 2003-2004)

	Systolic Blood Pressure (SBP)								
	CC analysis			IMI analysis			SSIMI analysis		
	Est.	s.e.	p-value	Est.	s.e.	p-value	Est.	s.e.	p-value
Intercept	87.80	1.16	<.0001	89.28	1.06	<.0001	87.53	1.35	<.0001
HHINC* (100k dollars)	-0.84	0.97	0.3907	-0.84	1.11	0.4574	-0.88	0.94	0.3482
EDU (years)	-2.30	0.57	<.0001	-2.06	0.44	<.0001	-2.38	0.55	<.0001
AGE(years)	0.49	0.01	<.0001	0.50	0.01	<.0001	0.50	0.01	<.0001
Female	3.31	0.48	<.0001	2.78	0.44	<.0001	3.15	0.46	<.0001
BMI(kg/m^2)	0.46	0.04	<.0001	0.41	0.03	<.0001	0.47	0.04	<.0001

	Diastolic Blood Pressure (DBP)								
	CC analysis			IMI analysis			SSIMI analysis		
	Est.	s.e.	p-value	Estimate	s.e.	p-value	Est.	s.e.	p-value
Intercept	45.46	1.06	<.0001	46.94	1.00	<.0001	45.46	1.19	<.0001
HHINC (100k dollars)	2.97	0.89	0.0008	2.82	0.87	0.0026	2.83	0.97	0.0050
EDU (years)	4.86	0.52	<.0001	4.06	0.43	<.0001	4.95	0.52	<.0001
AGE(years)	0.12	0.01	<.0001	0.11	0.01	<.0001	0.11	0.01	<.0001
Female	1.81	0.44	<.0001	1.83	0.36	<.0001	1.86	0.42	<.0001
BMI(kg/m^2)	0.43	0.04	<.0001	0.40	0.03	<.0001	0.44	0.04	<.0001

*HHINC: household income, in dollars multiplied by 100,000.

CHAPTER 3

A Pseudo Bayesian Shrinkage Approach to Regression with Missing Covariates

ABSTRACT: We consider the regression of outcome *Y* on regressors *W* and *Z* with some values of *W* missing, when our main interest is the effect of *Z* on *Y*, controlling for *W*. Three common approaches to regression with missing covariates are (a) complete-case analysis (CC), which discards the incomplete cases, and (b) ignorable likelihood methods, which base inference on the likelihood based on the observed data, assuming the missing data are missing at random (Rubin, 1976), and (c) nonignorable modeling, which posits a joint distribution of the variables and missing data indicators. Another simple practical approach that has not received much theoretical attention is to drop the regressor variables containing missing values from the regression modeling (DV, for drop variables). DV does not lead to bias when either (a) the regression coefficient of *W* is zero or (b) *W* and *Z* are uncorrelated. We propose a pseudo-Bayesian approach for regression with missing covariates that compromises between the CC and DV estimates, exploiting information in the incomplete cases when the data support DV assumptions. We illustrate favorable properties of the method by simulation, and apply the proposed

method to a liver cancer study. Extension of the method to more than one missing covariates is also discussed.

Some key words: Complete-case analysis, drop variables analysis, Gibbs sampling, nonignorable modeling, shrinkage, variable selection.

3.1 Introduction

We consider multivariate regression with missing covariates, with data displayed in Figure 3.1. There is a set of outcomes *Y* and two sets of regressor variables *Z* and *W*, with *Z* and *Y* fully observed and *W* with missing values. Here we assume *W* is a single variable, though generalization to multivariate *W* is possible and discussed later. We denote by (z_i, w_i, y_i) the values of (*Z*, *W*, *Y*) for observation *i*, and by R_{w_i} the indicator for whether *W* is observed or missing. Our main interest concerns one or more of the coefficients of the regression of the regression of *Y* on *Z*, adjusting for *W*. The incomplete cases have very little information for the coefficient of *W* (Little, 1992), and since our focus is on exploiting information in the incomplete cases, we assume that this coefficient is not the main parameter of interest. This kind of data structure is common in health-related studies. For example, in a behavioral intervention trial, the treatment assignment variable is always observed, while other variables may be missing. In a study of the effect of lead exposure on academic scores, blood lead level is always observed but socioeconomic variables such as Income might have missing values.

Reviews of regression with missing data include Little (1993), Ibrahim et al. (1999), Ibrahim et al. (2002), Ibrahim et al. (2005), Chen et al. (2008). Three common approaches are:

(a) Complete-case analysis (CC), which discards the incomplete cases;

(b) Ignorable likelihood methods (IL), which base inference on the observed likelihood given a model for the distribution of *Y* and *W* given *Z* that does not include a distribution for the missing data mechanism; examples of IL methods

include ignorable maximum likelihood, and multiple imputation based on draws from the Bayesian predictive distribution;

(c) Nonignorable modeling (NIM), which derives inference from the likelihood function based on a joint distribution of the variables and the missing data indicators. Examples include generalized Tobit (Type II) model (Heckman 1976, Amemiya 1984) and pattern-mixture models (Little 1993, 1994, Little and Wang, 1996).

IL methods are valid under well-specified models when the missing data are missing at random, which in this context means that missingness of W can depend on Z and Y but not on W. We focus here on situations where missingness of W is thought to depend on the value of W, so that IL methods are biased. One possibility is to apply an NIM method, but such methods are vulnerable to misspecification of the missing data mechanism, and suffer from problems with identifying the parameters (see e.g. Little and Rubin, 2002, chapter 15). Also software for these methods is not widely available.

A simple alternative is to apply CC in this setting. This has the advantage of yielding valid inferences when missingness of W depends on the covariates (Z, W) but not on the outcomes Y (Little and Rubin 2002, Example 3.3). On the other hand, it discards information in the incomplete cases, which might be substantial if the fraction of cases with W missing is high.

Another simple approach, which has received less theoretical attention but we suspect is common in practice, is to simply drop the incomplete variable from the analysis (DV), and estimate the regression of Y on Z using all the cases. It is well known from regression theory with complete data that omitting a covariate yields valid inferences when: (1) The

omitted covariate has no effect on the outcome; or (2) the missing covariate is not associated with the fully-observed regressors. If neither of these conditions holds, then DV leads to biased estimates. If the above effects are nonzero but small, DV is still an attractive method, since it may be worth accepting a small amount of bias in the regression estimates in order to retain the information in the incomplete cases.

A pragmatic two-step approach is to apply CC first, and then switch to DV if the coefficient of W in the CC analysis is small, for example if it has a non-significant P-Value. This can be viewed as a simple case of variable selection with missing data, which is considered more generally in Rubin (1976a). However, this is an "all or nothing" approach, and in general basing inferences on a preliminary statistical test is known to be problematic. This article proposes a Bayesian data-driven compromise between CC and DV, based on a prior distribution that assigns some weight to both analyses.

The rest of the article is organized as follows. Section 3.2 presents a motivating example using data from two Eastern Cooperative Oncology Group clinical trials. Section 3.3 reviews properties of CC and DV, in a slightly more general regression setting. In Section 3.4, we propose a pseudo-Bayesian shrinkage method for regression with missing covariates, which compromises between CC and DV analysis, assigning more weight to DV when the assumptions of that analysis are empirically justified, and more weight to CC when they are not. Section 3.5 presents some simulations that demonstrate attractive properties of the proposed method, and in section 3.6 we apply the proposed method to a liver cancer data set. Extensions to more than one missing regressors are discussed in Section 3.7.

3.2 The motivating example: a liver cancer study

To motivate our methodology, we consider data of 191 patients from Eastern Cooperative Oncology Group clinical trials EST 2282 (Falkson et al., 1990) and EST 1286 (Falkson et al., 1995). This dataset has been widely used to illustrate different methods for handling incomplete covariates in regression analysis or generalized linear models (Ibrahim et al. 1999, Huang et al. 2005, Chen et al. 2007, Das et al. 2010).

We are primarily interested in the patient's status as he/she enters the trials. In particular, we are interested in how the number of the cancerous liver nodes (CNTs) is predicted by four baseline characteristics:

(1) body mass index (BMI, in kg/m^2);

(2) age (in years);

(3) jaundice (yes, no): the yellowish staining of the skin and the whites of the eye;

(4) time since diagnosis of the disease (TSD, in weeks).

The effects of BMI, age, and jaundice are of more interest to a physician because these could be potential risk factors for liver cancer, but TSD is an important covariate that needs to be adjusted for.

Like many other empirical studies, this dataset contains missing values. TSD is missing for 17 patients (8.9%) while other variables are fully observed. CC analysis suffers from inefficiency and potential bias if the missingness of TSD depends on the outcome. DV analysis uses all cases but makes a strong assumption that exclusion of TSD does not bias the estimates of the other regression coefficients. IL makes use of the partial information in the incomplete case but assumes the missing data are missing at

random (MAR; Rubin 1976b, Little and Rubin 2002). We propose a pseudo-Bayesian approach for this problem, which compromises between the CC and DV estimates.

Before describing the pseudo-Bayesian approach, we first review more precisely the assumptions underlying the CC and DV methods.

3.3 Complete case and drop variable analyses

In this section, we consider the data with the structure in Figure 3.2. Let $\{(z_i, w_i, y_i), i = 1,..n\}$ denote n independent observations on a (possibly multivariate) outcome variable *Y* and two sets of covariates, *Z* and *W*, where *Z*, *Y* are fully observed and *W* has missing values. Interest concerns the parameters ϕ of the distribution of *Y* given (*Z*, *W*), say $p(y_i \mid z_i, w_i, \phi)$.

The rows of Figure 3.2 divide the cases into two patterns. Pattern 1 (i = 1,…,m) consists of complete cases, for which (z_i, w_i, y_i) are fully observed. Pattern 2 consists of cases where at least one of the variables in w_i is missing. The column R_{w_i} represents a vector of response indicators for w_i, with entries 1 if a variable is observed and 0 if a variable is missing. For the complete cases, $R_{w_i} = u_w \equiv (1,...,1)$, a vector of ones of the same length as w_i, indicating that all the entries in w_i are observed. For the incomplete cases in Pattern 2, we write $\bar{u}_w$, defined to mean that some entries in R_{w_i} are 0 and others are 1. The pattern of missing values will typically vary over the individual rows in Pattern 2, but we do not need to distinguish them for the present discussion.

Our main interest is the effect of *Z* on *Y*, adjusting for *W*. CC analysis bases inferences for ϕ on the complete observations in Pattern 1. In a likelihood setting, the method bases inference on the conditional likelihood corresponding to the complete cases, namely:

$$L_{\mathrm{cc}}(\phi) = \text{const.} \times \prod_{i=1}^{m} p\ \ y_i \mid w_i, z_i, R_{w_i} = u_w; \phi\ \ , \tag{3.1}$$

The key condition under which inference based on $L_{\mathrm{cc}}(\phi)$ is valid is that the probability that an observation is complete does not depend on the outcomes, that is:

$$p\ \ R_{w_i} = u_w \mid z_i, w_i, y_i, \psi)\ \ = p\ \ R_{w_i} = u_w \mid z_i, w_i, \psi)\ \ \text{for all } y_i \tag{3.2}$$

Note that this condition is missing not at random (MNAR), since missingness depends on the values of *W* which are sometimes missing. CC analysis works in this case because Eq. (2) implies that

$$p\ \ y_i \mid w_i, z_i, R_{w_i} = u_w, \phi\ \ = p\ \ y_i \mid w_i, z_i, \phi\ \ ,$$

so the regression based on the complete cases is the regression of interest, for the whole sample. Technically, inference based on (1) can be considered a partial likelihood method (Little and Zhang, 2011). The likelihood for a fully specified model with parameters (ϕ, ψ) can be written as

$$L(\phi, \psi \mid Z, W_{\mathrm{obs}}, Y_{\mathrm{obs}}, R_w) = L_{\mathrm{cc}}(\phi) L_{\mathrm{rest}}(\phi, \psi)\ ,$$

and the component $L_{\mathrm{rest}}(\phi, \psi)$ is discarded. ML estimates based on $L_{\mathrm{cc}}(\phi)$ are consistent and asymptotically normal, but are not necessarily fully efficient, since $L_{\mathrm{rest}}(\phi, \psi)$ may contain information about the parameters of interest ϕ. However, recovering this

information requires a model for the missing data mechanism, which may be difficult to specify correctly, and which is not needed for CC analysis.

Instead of dropping the incomplete cases, DV analysis removes the incomplete variable from the regression model, as would be sensible if β_w, the regression coefficient of *W*, were equal to zero. Writing $\phi = \beta_w, \phi_z$, the method bases inference on the following likelihood:

$$L_{\mathrm{DV}}(\phi_z) = \text{const.} \times \prod_{i=1}^{n} p\ \ y_i \mid z_i, \beta_w = (0,..,0); \phi_z\ \ , \tag{3.3}$$

When *W* has no effect on the outcome *Y*, DV analysis is better than CC, not only because it removes inefficiency induced by estimating the coefficient of *W*, but also by retaining the incomplete cases. The DV analysis also yields valid inferences for the regression coefficient of *Z* even if $\beta_w \neq (0,..,0)$ when *W* and *Z* are not associated. This fact will be exploited in the proposed method, which we now describe.

3.4 Pseudo-Bayesian Shrinkage Method for Regression with Missing Covariates

3.4.1 Motivation

In this section, we consider the data structure in Figure 3.1, where the missing covariate *W* is univariate and the fully observed *Z* could be multivariate. We are interested in the regression of *Y* on *Z*, controlling for *W*, and assume the normal linear regression model:

$$y_i \mid w_i, z_i, \beta_w, \beta_z, \sigma^2\ \ \sim N\ \ \beta_0 + w_i\beta_w + z_i\beta_z^T; \sigma^2\ \ ,\ i = 1,...,n.$$

The CC analysis is valid when the missingness of *W* does not depend on the outcome *Y*, after conditioning on *Z* and *W*. DV analysis is valid if either of the following two conditions is met:

(I) $\beta_w = 0$;

(II). $\rho_{wz} \equiv \text{cov}(W, Z^*) = 0$, where *Z** is a linear combination of individual components of *Z*, with the weights being the corresponding estimated regression coefficients in the regression of *Y* on *W* and *Z*.

This suggests assigning β_w a prior distribution that assigns positive probability to 0, since this will recover information in the incomplete cases when the posterior probability that $\beta_w = 0$ is high. This kind of prior has been proposed for Bayesian variable selection problems. One example is the 'spike and slab' mixture prior, which puts a probability mass on $\beta_w = 0$ (Mitchell and Beauchamp 1988). Another example is using a mixture of two normal distributions with zero mean and different variances, a formulation proposed by George and McCulloch (1993). In this article, we model β_w using mixture of a point mass at $\beta_w = 0$ and a normal distribution with zero mean and large variance.

3.4.2 Modeling

Introducing a latent variable $J(= 0 \text{ or } 1)$, we represent the mixture distribution by

$$\beta_w \mid J \sim J\delta\ 0\ +\ 1-J\ \ N\ 0, \tau_w^2 \qquad (3.4)$$

with $\delta\ 0$ representing a point mass at 0, and

$$\Pr\ J=0\ =1-\Pr\ J=1\ =\pi_0 \qquad (3.5)$$

When $J = 0$, $\beta_w \sim N\ 0, \tau_w^2$, and when $J = 1$, $\beta_w \equiv 0$. We set τ_w^2 large so that if $J = 0$, β_w has a flat prior as in a standard least squares analysis. To incorporate (3.4) in the full prior distribution, we use a multivariate normal prior

$$\beta \mid J \sim N\ 0, D_J D_J \tag{3.6}$$

with

$$D_J \equiv diag\ 1, a\tau_w, \tau_z\ , \tag{3.7}$$

with $a = 1$ if $J = 0$ and $a = 0$ if $J = 1$.

We use the inverse gamma conjugate prior for the residual variance σ^2,

$$\sigma^2 \mid J \sim IG\ \nu_J / 2, \nu_J \lambda_J / 2\ . \tag{3.8}$$

The choices of ν_J and λ_J reflects the statistician's prior belief about the residual variances for whether the covariate *W* is included in the model or not. In the absence of such prior information, we choose ν_J and λ_J small so that the analyses are mainly based on the likelihood.

3.4.3 A pragmatic choice of π_0

As we can see from section 3.4.1, one condition for DV analysis to be valid is that the correlation coefficient ρ_{wz} between *W* and *Z** is zero. This indicates that: (1) if we believe that $\rho_{wz} = 0$, then we can put a high prior probability on Pr $J = 1$, and (2) on the other hand, if $|\rho_{wz}|$ is large, we are more inclined to include *W* and use the complete-case analysis. So from a pragmatic perspective, it is advantageous to choose π_0 as an

increasing function of $|\rho_{wz}|$. We found the following choice to work well in simulation studies:

$$\pi_0 = f\ \ |\rho_{wz}|\ \ = |\rho_{wz}|. \tag{3.9}$$

To propagate the variation in posterior estimation of β_z, we recommend using draws of ρ_{wz} based on Bayesian predictive distribution based on model *W* and *Z** as a bivariate normal distribution using the complete-cases likelihood

$$L\ \ \mu_{wz}, \Sigma_{wz}\ \ = \prod_{i=1}^{m} f\ \ w_i, z_i \mid \mu_{wz}, \Sigma_{wz} \tag{3.10}$$

3.4.4 Estimation

We obtain draws of the parameters from the posterior distribution using the following Gibbs-like sampler. Let

$$\beta = \ 1, \beta_w, \beta_z^T\ ^T,\ X_{CC} = \begin{bmatrix} 1 & w_1 & z_1^T \\ 1 & w_2 & z_2^T \\ & .... & \\ 1 & w_m & z_m^T \end{bmatrix},\ X_{DV} = \begin{bmatrix} 1 & z_1^T \\ 1 & z_2^T \\ & .... \\ 1 & z_n^T \end{bmatrix},\ Y_{CC} = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_m \end{bmatrix},\ Y_{DV} = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix}.$$

Also, let $\hat{\beta}_{\text{DV,LS}}$ be the least square estimate based on DV analysis, and $\hat{\beta}_{\text{CC,LS}}$ be the least square estimate based on complete-case analysis. Accordingly, the estimated residual variances for DV and CC analysis are denoted as $\hat{\sigma}_{\text{DV,LS}}$ and $\hat{\sigma}_{\text{CC,LS}}$.

The chain is initialized at a starting value β^0, σ^0, J^0. A reasonable starting value for J^0 is 0, which is a complete-case scenario, and therefore the corresponding starting values for β^0, σ^0 are $\hat{\beta}_{\text{CC,LS}}$ and $\hat{\sigma}_{\text{CC,LS}}$.First, β^k, σ^k can be sampled in the following way:

(1) If $J^{k-1} = 0$,

$$\beta^k = \left(\beta_0^k, \beta_w^k, \beta_z^k\right)^T \sim N\left(A_{\mathrm{CC},J^{k-1}}\left(\sigma^{k-1}\right)^{-2} X_{\mathrm{CC}}^T X_{\mathrm{CC}} \hat{\beta}_{\mathrm{CC,LS}}, A_{\mathrm{CC},J^{k-1}}\right), \tag{3.11}$$

where $A_{\mathrm{CC},J^{k-1}} = \left(\left(\sigma^{k-1}\right)^{-2} X_{\mathrm{CC}}^T X_{\mathrm{CC}} + D_{J^{k-1}}^{-1} D_{J^{k-1}}^{-1}\right)^{-1}$ and $D_J^{-1} = diag\left[1, \tau_w^{-1}, \tau_z^{-1}\right]$;

and $\sigma^{2(k)}$ is obtained by sampling from

$$\sigma^{2(k)} \sim f\left(\sigma^{2(k)} \mid \beta^k, J^{k-1}\right) = IG\left(\frac{n+\nu_{J^{k-1}}}{2}, \frac{\left|Y_{\mathrm{CC}} - X_{\mathrm{CC}}\beta^k\right|^2 + \nu_{J^{k-1}}\lambda_{J^{k-1}}}{2}\right). \tag{3.12}$$

(2) If $J^{k-1} = 1$,

$\beta^k = \left(\beta_0^k, 0, \beta_z^k\right)^T$ with

$$\left(\beta_0^k, \beta_z^k\right)^T \sim N\left(A_{DV,J^{k-1}}\left(\sigma^{k-1}\right)^{-2} X_{\mathrm{DV}}^T X_{\mathrm{DV}} \hat{\beta}_{\mathrm{DV,LS}}, A_{\mathrm{DV},J^{k-1}}\right), \tag{3.13}$$

where $A_{\mathrm{DV},J^{k-1}} = \left(\left(\sigma^{k-1}\right)^{-2} X_{\mathrm{DV}}^T X_{\mathrm{DV}} + D_{J^{k-1}}^{-1} D_{J^{k-1}}^{-1}\right)^{-1}$ and $D_J^{-1} = diag\left[1 \;\; \tau_z^{-1}\right]$;

and $\sigma^{2(k)}$ is obtained by sampling from

$$\sigma^{2(k)} \sim f\left(\sigma^{2(k)} \mid \beta^k, J^{k-1}\right) = IG\left(\frac{n+\nu_{J^{k-1}}}{2}, \frac{\left|Y_{\mathrm{DV}} - X_{\mathrm{DV}}\left(\beta_0, \beta_z^k\right)^T\right|^2 + \nu_{J^{k-1}}\lambda_{J^{k-1}}}{2}\right). \tag{3.14}$$

Next, ρ_{zw}^k is sampled based on the posterior covariance matrix of the bivariate normal distribution formed by *Z** and *W* (using the complete-cases).

The final step is to sample J^k, which is Bernoulli with probability

$$\Pr\left(J^k = 1 \mid \beta^k, \sigma^j, \rho_{zs}^k\right) = \frac{r}{r+s}, \tag{3.15}$$

with $r = f\ \ \beta^k \mid J^k = 1\ \ 1 - \left|\rho_{zw}^k\right|$ and $s = f\ \ \beta^k \mid J^k = 0\ \ \left|\rho_{zw}^k\right|$.

Note that, when $J^{k-1} = 0$, the conditional distribution of β^k, σ^k are based on the complete-case likelihood, which is a partial likelihood. Since partial likelihood is not very principled from a strict Bayesian perspective, we label the method "pseudo-Bayes". We demonstrate in simulations in the next section that it leads to inferences with good frequentist properties.

3.4.5 Posterior probability that $J = 0$, π_1

The posterior probability of $J = 1$, π_1, namely, the posterior probability of including the incomplete variable *W* in the regression model and using complete case analysis, is an important indicator in the modeling. A small π_1 tends to put more weight on DV, whereas a large π_1 puts more weight on CC.

3.5 Simulation studies

In this section we describe simulations that illustrate the properties of the pseudo-Bayesian approach in Section 3.4.

We simulate $w, z_1, z_2\ _i$ from normal distribution with mean 0, and covariance matrix

$$\begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

for $i = 1, 2, ..., 100.$ *Y* is related to *Z* and *W* by the linear model

$$y_i = 1 + aw_i + z_{1i} + z_{2i} + \varepsilon_i,\ \ \varepsilon_i \overset{iid}{\sim} N\ \ 0, 2.5^2\ \ .$$

Let M_{w_i} denote the missing data indicator for w_i. Missing values in *W* are generated based on the following five missing data mechanisms:

(I) MCAR: $\Pr\ M_{w_i} = 1 | w_i, z_{1i}, z_{2i}, y_i\ = 0.25$;

(II) MNAR: $\Pr\ M_{w_i} = 1 | w_i, z_{1i}, z_{2i}, y_i\ = \text{expit}\ w_i - 1$;

(III) MAR: $\Pr\ M_{w_i} = 1 | w_i, z_{1i}, z_{2i}, y_i\ = \text{expit}\ z_{1i} + z_{2i} - 1$;

(IV) MNAR2: $\Pr\ M_{w_i} = 1 | w_i, z_{1i}, z_{2i}, y_i\ = \text{expit}\ w_i + z_{1i} + z_{2i} - 1$;

(V) MAR2: $\Pr\ M_{w_i} = 1 | w_i, z_{1i}, z_{2i}, y_i\ = \text{expit}\ y_i - 2$.

where $\text{expit}\ \cdot$ is inverse logit function, $\text{expit}\ \cdot = \exp\ \cdot\ /\ 1 + \exp\ \cdot$. Each missing data generation scheme results in about 25% of the values of *W* being missing.

We simulate data for three different correlation coefficients (ρ =0, 0.3, 0.8) and two regression coefficients for *W* (a = 0, 1), yielding 30 scenarios.

Five methods are applied to estimate the regression coefficients:

(1) BD: estimates from the regression before deletion (BD), as a benchmark method.

(2) IL: ignorable maximum likelihood method assuming MAR;

(3) CC: Complete-case analysis;

(4) DV: dropping the missing covariate *W*;

(5) PB: pseudo-Bayesian shrinkage method between CC and DV;

We report the ratios of RMSEs of IL, CC, DV and PB to the RMSE of BD, confidence coverage probabilities and empirical bias z-score (which is calculated using empirical bias/empirical standard error of the mean) of the estimated regression

coefficients from each method, in Tables 3.1, 3.2 and 3.3. Results are based on 1000 repetitions for each simulation condition. Table 3.1 also reports the posterior probabilities of including *W*.

We focus on the regression coefficients of z_1 and z_2. CC is consistent for the first four missing data mechanisms since missingness does not depend on the outcome, but biased for the fifth missing data mechanism since missingness of *W* is dependent on the outcome. There is some loss of information since the incomplete cases are dropped from the analysis. IL is consistent and efficient for missing data mechanism I, III and V since all are missing at random. DV is valid when a =0 or $\rho = 0$, and in these cases the pseudo-Bayesian (PB) estimates are close to DV; when $a \neq 0$ and $\rho \neq 0$, PB yields a compromise between CC and DV, with the posterior probability assigned to CC estimates increasing as a and/or ρ move away from 0. The method yields small RMSEs and good confidence coverage compared to CC and DV in almost all scenarios. As expected, IL performs well for the missing at random mechanisms I, III and IV but exhibits some bias when the data are not missing at random.

3.6. Application to a liver cancer study

We now apply the proposed method to the liver cancer data presented in Section 3.2. We regress the baseline number of cancerous liver nodes (CNTs) on four baseline charactistics: body mass index (BMI), age in year, associated jaundice (yes, no) and time since diagnosis of the disease (TSD, in weeks). To be consistent with Chen, Zeng, and Ibrahim (2007), we use the same transformation as they did. Square root transformations are used on CNTs and TSD to achieve approximate normality. The new continuous

explanatory variables BMI, Age and $\sqrt{TSD}$ are then formed by dividing the original variables by 50, 70, and 18 respectively, to bound the covariates on the interval of (0, 1). In Chen, Zeng and Ibrahim (2007), TSD is assumed to be missing at random. However, it is likely that TSD is not MAR since patients with longer TSD are less likely to recall the date of diagnosis of liver cancer, which means missingness of TSD depends on TSD itself.

The Pearson correlation between TSD and BMI, Age, Jaundice are -.020, .013, and .009 respectively. The correlation between TSD and the linear combination of BMI, Age, and Jaundice weighted by the regression coefficients using complete cases is -.002.

Table 3.4 shows the results of applying the pseudo-Bayesian shrinkage method, CC and DV. We run 10000 iterations and obtain draws of the posterior estimates of the regression coefficients. The posterior probability of including TSD is 0.0153, indicating that the pseudo Bayesian method favors dropping TSD from the regression and using full sample. This is not surprising, since the correlation between TSD and other covariates is small, and the effect of TSD on the outcome CNTs is also small.

For easier comparison, we calculate a pseudo p-value based on t-distribution. The degree of freedom is calculated using the following formula:

$$df = n * (1 - \hat{\pi}_1) - k - \hat{\pi}_1 \tag{3.16}$$

where n is the full sample size, k is number of all regressors, and $\hat{\pi}_1$ is the estimated posterior probability of including the missing regressor.

As we can see from Table 3.4, both complete-case analysis and pseudo-Bayesian method show that Age and Jaundice are related to the number of cancerous liver nodes,

while BMI and TSD are not significant. However, the pseudo-Bayesian method yields smaller standard error for the regression estimate, so the effect of Age and Jaundice are stronger than complete case analysis. Since the posterior probability of keeping TSD in the modeling is very small ($\hat{\pi}_1 = 0.0153$), the pseudo Bayesian method is very similar to the regression without TSD.

3.7. Discussion

We have described a pseudo-Bayesian shrinkage method for regression analysis with a missing covariate, which is a compromise between complete-case analysis and the analysis that drops the missing covariate. The method recovers information in the incomplete cases by assigning the regression coefficient of the incomplete variable a "slab and spike" prior with positive prior probability of being zero. A Gibbs-like iterative sampling algorithm is used to implement the method; convergence is fast.

The method is appropriate when missingness of the missing covariate depends on the covariates but not the outcome. This mechanism is potentially missing not at random, and an attraction of the proposed method is that it handles such cases without having to model the specific form of the missing data mechanism. The method also works when the missing data mechanism for the covariate is MAR but independent of the outcome. However in general ignorable likelihood methods are preferable in that case, since they are asymptotically efficient.

Our method can be generalized to the situation when w_i is a vector with dimension *d*, with components missing on possibly different sets of cases. We assume

that the missingness of *W* is independent of the outcome. We assign to each component of *W* an independent mixture distribution prior as in Section 3.4.2. In this case, $\rho_{w_1 z},...,\rho_{w_d z}$ represents the *d* correlation coefficients between *W* and *Z**, and draws of the indicators $J_1,...,J_d$ for whether the corresponding coefficients are zero are sampled in the estimation step. We suggest sampling $J_1,...,J_d$ in a random order to get fast convergence of the chain. For the *j*th component of *W*, $Y_{\mathrm{CC},j}$ is defined to be the vector of outcomes corresponding to complete-case analysis, while $Y_{\mathrm{DV},j}$ is defined to be the vector of outcomes corresponding to the complete cases when W_j is dropped from the regression. $X_{\mathrm{CC},j}$, $X_{\mathrm{DV},j}$, $\hat{\beta}_{\mathrm{CC},j}$ and $\hat{\beta}_{\mathrm{DV},j}$ are defined in a similar fashion.

The proposed method could be combined with existing multiple imputation methods to handle more general problems where *Z* is also incomplete. In particular, when missingness of covariates *W* is MNAR but does not depend on the outcome, and missingness of *Z* is MAR, the method could also be applied by assigning similar mixture priors to the regression coefficients of *W*, while using multiple imputation via chained equations (Raghunathan et al., 2001; IVEware, 2011; MICE, 2011) to impute missing values of *Z*.

There is a potential loss of efficiency of the pseudo-Bayesian approach compared to full modeling of the data and missing-data mechanism. However, the proposed method avoids specifying a model for the missing data indicators, which is vulnerable to model misspecification. Future work will examine this trade-off in more detail.

Figure 3.1 Missing Data Structure in Section 3.1

Pattern	Observation, i	z_i	w_i	y_i	R_{w_i}
1	$i = 1,\ldots,m$	√	√	√	1
2	$i = m+1,\ldots,n$	√	x	√	0

Key: √ denotes observed, x denotes missing

Figure 3.2 Data Structure for Section 3.3

Pattern	Observation, i	z_i	w_i	y_i	R_{w_i}
1	$i = 1,\ldots,m$	√	√	√	$u_w = (1,\ldots,1)$
2	$i = m+1,\ldots,n$	√	?	√	$\overline{u}_w$

Key: √ denotes observed, ? denotes missing at least one entry

Table 3.1: RMSE Ratios for Individual Regression Coefficients under Five Missing Data Mechanisms (1000 replications)

ρ	a	MD Mechanism	$\hat{\pi}_1$	β_0				β_w				β_{z1}				β_{z2}			
				IL	CC	DV	PB	IL	CC	DV	PB	IL	CC	DV	PB	IL	CC	DV	PB
0	0	MCAR	0.040	1.00	1.14	0.99	0.99	1.15	1.14	0.00	0.06	1.00	1.17	1.00	0.99	1.00	1.19	0.99	0.98
		MNAR	0.042	1.06	1.25	1.00	1.00	1.29	1.26	0.00	0.08	1.01	1.19	1.00	0.99	1.01	1.22	1.00	0.99
		MAR	0.043	1.02	1.31	1.00	0.99	1.25	1.22	0.00	0.07	1.01	1.31	0.99	0.98	1.01	1.31	0.99	0.99
		MNAR2	0.063	1.09	1.36	0.99	1.00	1.40	1.37	0.00	0.13	1.03	1.35	1.00	0.99	1.04	1.36	1.00	1.00
		MAR2	0.044	1.04	5.12	0.99	1.01	1.60	1.02	0.00	0.06	1.03	1.77	1.00	0.98	1.01	1.75	0.99	0.98
	1	MCAR	0.126	1.03	1.19	1.07	1.06	1.13	1.17	3.85	3.32	1.04	1.19	1.08	1.07	1.02	1.17	1.08	1.05
		MNAR	0.124	1.42	1.22	1.07	1.08	1.26	1.26	3.80	3.32	1.04	1.25	1.06	1.04	1.04	1.24	1.05	1.03
		MAR	0.127	1.05	1.29	1.06	1.06	1.20	1.23	3.90	3.38	1.05	1.34	1.06	1.05	1.03	1.31	1.06	1.04
		MNAR2	0.167	1.47	1.40	1.07	1.07	1.31	1.33	3.84	3.21	1.16	1.41	1.07	1.06	1.14	1.33	1.06	1.05
		MAR2	0.092	1.13	5.36	1.07	1.15	1.52	1.76	4.02	3.75	1.07	1.81	1.09	1.07	1.07	1.64	1.05	1.03
0.3	0	MCAR	0.139	1.00	1.14	1.00	1.00	1.22	1.20	0.00	0.23	1.02	1.17	0.97	0.97	1.02	1.13	0.98	0.96
		MNAR	0.132	1.05	1.24	0.99	1.00	1.32	1.30	0.00	0.23	1.02	1.24	0.97	0.98	1.02	1.22	0.96	0.96
		MAR	0.120	1.01	1.37	1.00	1.00	1.25	1.22	0.00	0.21	1.02	1.33	0.97	0.98	1.02	1.28	0.96	0.96
		MNAR2	0.074	1.08	1.47	1.00	0.99	1.33	1.30	0.00	0.14	0.99	1.35	0.95	0.95	1.00	1.30	0.96	0.96
		MAR2	0.127	1.03	4.92	1.00	1.23	1.65	1.08	0.00	0.19	1.03	1.64	0.97	0.95	1.05	1.54	0.96	0.93
	1	MCAR	0.340	1.03	1.17	1.06	1.04	1.14	1.15	3.46	2.36	1.02	1.15	1.34	1.13	1.04	1.19	1.32	1.12
		MNAR	0.277	1.35	1.24	1.06	1.04	1.32	1.34	3.65	2.69	1.06	1.26	1.36	1.19	1.05	1.24	1.26	1.13
		MAR	0.276	1.05	1.33	1.06	1.06	1.25	1.28	3.72	2.74	1.05	1.27	1.39	1.20	1.08	1.33	1.29	1.16
		MNAR2	0.174	1.40	1.48	1.05	1.10	1.32	1.33	3.59	2.99	1.09	1.33	1.38	1.24	1.08	1.32	1.29	1.17
		MAR2	0.169	1.09	5.34	1.03	1.41	1.44	1.60	3.78	3.32	1.09	1.61	1.33	1.09	1.11	1.70	1.37	1.12
0.8	0	MCAR	0.598	1.00	1.16	0.99	1.06	1.18	1.17	0.00	0.77	1.04	1.18	0.90	1.00	1.03	1.17	0.88	0.99
		MNAR	0.572	1.04	1.28	1.00	1.11	1.27	1.24	0.00	0.80	1.02	1.19	0.87	0.98	1.08	1.28	0.89	1.03
		MAR	0.530	1.02	1.52	1.00	1.10	1.31	1.28	0.00	0.77	1.07	1.27	0.89	1.00	1.07	1.25	0.90	0.99
		MNAR2	0.475	1.02	1.56	1.00	1.13	1.34	1.31	0.00	0.74	1.05	1.26	0.89	0.96	1.06	1.30	0.90	0.98
		MAR2	0.555	1.04	5.30	1.00	3.05	1.66	1.12	0.00	0.72	1.15	1.19	0.88	0.91	1.16	1.29	0.90	0.98
	1	MCAR	0.714	1.02	1.17	1.02	1.10	1.20	1.18	2.14	1.18	1.05	1.20	1.30	1.10	1.05	1.17	1.34	1.09
		MNAR	0.675	1.07	1.23	1.02	1.14	1.26	1.25	2.13	1.24	1.06	1.24	1.34	1.13	1.06	1.21	1.29	1.10
		MAR	0.643	1.04	1.41	1.00	1.19	1.29	1.28	2.13	1.28	1.09	1.30	1.36	1.15	1.07	1.27	1.25	1.11
		MNAR2	0.578	1.12	1.62	1.02	1.23	1.35	1.33	2.10	1.34	1.07	1.30	1.26	1.11	1.10	1.39	1.37	1.20
		MAR2	0.592	1.06	5.32	1.01	3.22	1.48	1.24	1.99	1.37	1.16	1.28	1.24	0.98	1.17	1.31	1.34	1.00

Table 3.2: 95% Confidence Coverage for Individual Regression Coefficients under Five Missing Data Mechanisms

ρ	a	MD Mechanism	β_0						β_w						β_{z1}						β_{z2}				
			BD	IL	CC	DV	PB		BD	IL	CC	DV	PB		BD	IL	CC	DV	PB		BD	IL	CC	DV	PB
0	0	MCAR	94.0	93.9	95.1	94.1	94.3		93.8	94.0	95.1	100.0	100.0		94.2	94.5	94.7	93.4	94.9		93.2	93.2	92.7	93.4	93.9
		MNAR	95.3	95.4	95.0	95.6	95.6		94.2	93.9	94.7	100.0	100.0		93.2	93.0	94.4	93.5	93.5		93.2	93.7	93.4	93.5	94.5
		MAR	92.9	93.8	94.6	93.4	93.7		95.1	94.8	95.8	100.0	100.0		94.1	94.0	94.0	93.7	95.4		93.3	93.6	94.6	93.7	94.5
		MNAR2	93.8	93.8	93.6	94.4	94.1		94.7	93.1	93.4	100	100		94.7	94.1	94.2	95.3	95.7		95.2	94.0	94.7	95.3	95.4
		MAR2	94.6	95.3	0.4	94.6	89.8		94.9	92.1	94.5	100.0	100.0		95.7	95.5	74.6	94.9	96.2		94.8	95.4	73.0	94.9	95.6
	1	MCAR	94.6	94.8	95.5	95.5	92.9		94.0	94.1	94.4	0.0	65.9		95.5	93.5	94.6	94.9	95.9		94.8	94.8	95.0	93.7	94.9
		MNAR	94.9	86.3	95.5	95.5	95.4		94.8	94.6	94.7	0.0	61.3		94.3	94.1	94.2	94.9	95.9		95.2	94.4	94.0	95.7	96.2
		MAR	93.8	94.3	94.2	94.7	94.7		93.8	94.5	94.2	0.0	62.0		94.5	95.3	94.4	95.0	96.8		95.8	95.2	93.3	95.5	96.8
		MNAR2	96.2	87.3	94.5	95.3	95.0		94.6	93.5	94.1	0.0	68		95.1	92.9	93.9	96.1	97.1		96.0	93.6	94.4	96.0	97.1
		MAR2	95.7	95.7	0.3	95.4	59.9		95.7	93.4	76.6	0.0	26.4		96.0	95.7	76.2	94.4	96.1		94.7	94.3	77.7	95.8	97.1
0.3	0	MCAR	93.7	92.9	93.5	93.6	93.6		95.3	93.2	94.1	100.0	100.0		94.7	94.1	94.7	94.9	95.1		95.3	94.9	95.6	94.9	95.8
		MNAR	93.9	93.8	93.5	94.2	94.1		95.0	94.5	95.1	100.0	100.0		95.5	95.3	95.3	95.4	96.3		95.4	93.9	94.4	95.4	96.2
		MAR	94.4	93.6	93.6	94.5	94.0		95.2	94.1	94.2	100.0	100.0		93.4	94.4	94.1	94.6	95.4		93.9	94.6	94.8	94.6	95.4
		MNAR2	94.9	95.9	95.8	95.1	95.3		93.1	93.2	93.3	100.0	100		93.6	93.8	92.7	94.2	95.2		95.1	94.5	95.0	94.2	95.8
		MAR2	93.5	94.5	0.2	93.6	87.7		94.3	91.4	93.8	100.0	100.0		94.8	95.7	77.1	93.7	96.3		93.1	93.1	77.8	93.7	96.4
	1	MCAR	95.3	94.8	95.0	95.5	96.7		94.2	93.0	94.0	0.0	85.3		95.0	93.8	94.3	85.0	95.1		95.0	94.5	95.1	88.1	95.6
		MNAR	93.8	86.4	94.8	94.2	94.3		94.4	94.3	95.0	0.0	81.5		95.4	94.4	95.6	86.1	94.7		94.3	92.9	93.7	87.8	94.8
		MAR	94.3	93.4	94.2	94.7	95.4		95.4	94.0	94.4	0.0	83.2		94.8	95.1	95.4	85.8	95.6		94.4	93.5	93.8	86.8	95.2
		MNAR2	93.5	86.1	92.5	94.3	94.2		95.3	94.1	94.3	0.0	69		95.0	93.6	94.0	83.6	93.1		94.7	93.9	94.9	87.6	94.4
		MAR2	94.9	95.5	0.5	95.7	58.1		95.6	95.0	82.8	0.0	43.3		95.4	94.7	81.1	87.0	98.6		95.0	95.4	80.2	87.6	98.4
0.8	0	MCAR	93.7	94.5	93.9	93.7	94.0		95.3	94.7	95.3	100.0	100.0		94.8	95.2	94.7	94.4	96.1		94.2	94.4	94.4	94.4	96.6
		MNAR	94.3	94.3	93.1	94.1	93.3		93.0	92.3	92.6	100.0	100.0		94.3	94.6	94.3	95.2	96.4		95.0	95.9	94.2	95.2	96.8
		MAR	96.2	96.5	94.7	95.7	95.0		95.4	94.6	95.2	100.0	100.0		94.9	93.9	94.2	94.5	96.9		95.4	95.6	95.8	94.5	98.1
		MNAR2	94.1	94.3	93.1	93.9	93.3		94.5	94.3	94.7	100.0	100		95.4	95.5	95.0	95.4	97.8		95.3	94.5	95.3	95.4	98.0
		MAR2	94.3	96.0	0.8	94.9	77.3		94.3	92.3	93.5	100.0	100.0		93.8	94.2	90.6	93.6	97.2		93.9	94.0	88.5	93.6	96.9
	1	MCAR	95.7	96.4	95.5	95.2	95.5		94.6	95.6	94.6	0.0	95.1		94.5	95.8	93.8	82.4	95.9		94.0	95.0	93.9	81.2	95.1
		MNAR	94.9	94.0	95.1	94.3	94.6		95.6	94.6	96.0	0.0	95.6		95.5	95.4	93.3	81.0	96.1		93.9	94.2	94.0	83.0	96.6
		MAR	95.2	96.2	94.6	95.3	94.8		94.9	93.5	94.5	0.0	94.4		94.2	92.7	93.3	79.3	94.8		94.4	94.0	93.0	82.4	95.9
		MNAR2	96.4	94.6	94.9	95.5	95.1		94.6	92.6	94.0	0.0	92		94.1	94.8	95.3	83.3	97.5		95.2	94.9	93.4	81.2	95.7
		MAR2	95.1	94.6	1.6	95.3	57.1		93.4	93.2	89.4	0.0	85.2		94.0	93.1	88.2	82.5	98.5		94.8	94.3	91.8	82.0	98.6

Table 3.3: Bias (Z-score) for Individual Regression Coefficients under Five Missing Data Mechanisms (1000 replications)

ρ	a	MD Mechanism	β_0				β_w				β_{z1}				β_{z2}			
			IL	CC	DV	PB	IL	CC	DV	PB	IL	CC	DV	PB	IL	CC	DV	PB
0	0	MCAR	0.00	-0.01	0.00	-0.02	0.04	0.04	-	0.04	0.01	0.00	0.01	-0.03	0.01	0.00	0.02	-0.02
		MNAR	0.07	0.07	0.08	0.04	-0.05	-0.05	-	-0.03	0.04	0.03	0.03	0.00	0.02	-0.01	0.02	-0.02
		MAR	0.04	0.03	0.03	0.04	-0.03	-0.03	-	-0.02	0.00	-0.01	0.00	-0.05	0.00	0.02	-0.01	-0.05
		MNAR2	0.04	0.00	0.02	0.02	0.02	0.02	-	0.01	0.04	0.02	0.04	0.00	-0.01	-0.03	-0.02	-0.06
		MAR2	-0.10	-4.85	0.04	-0.25	0.00	-0.01	-	-0.01	-0.04	-1.25	-0.01	-0.11	-0.08	-1.31	-0.05	-0.15
	1	MCAR	0.02	0.05	0.02	0.03	0.05	0.03	-	-4.69	-0.06	-0.06	-0.07	-0.10	0.08	0.08	0.05	0.01
		MNAR	0.92	0.06	0.07	0.03	0.08	-0.01	-	-5.47	-0.01	-0.01	-0.01	-0.06	0.05	0.05	0.04	0.01
		MAR	0.01	0.01	-0.01	0.01	0.02	-0.05	-	-4.97	0.01	0.01	0.00	-0.04	-0.02	0.00	-0.04	-0.08
		MNAR2	0.86	-0.04	-0.04	-0.02	0.09	0.00	-	-3.67	0.39	-0.01	0.02	0.00	0.36	-0.05	-0.02	-0.04
		MAR2	-0.07	-4.86	0.03	-0.41	0.02	-1.19	-	-11.01	-0.08	-1.22	-0.07	-0.21	-0.02	-1.17	-0.02	-0.16
0.3	0	MCAR	-0.04	-0.03	-0.04	0.00	-0.03	-0.04	-	-0.01	0.02	0.05	0.01	-0.01	0.03	0.01	0.02	-0.01
		MNAR	0.00	0.02	-0.01	-0.02	0.03	0.03	-	0.03	-0.07	-0.08	-0.06	-0.09	-0.02	-0.01	-0.01	-0.05
		MAR	-0.04	-0.05	-0.04	-0.04	-0.01	0.00	-	0.02	0.01	-0.03	0.00	-0.03	-0.03	-0.02	-0.03	-0.07
		MNAR2	0.05	0.05	0.06	-0.01	0.00	0.00	-	0.00	0.03	0.03	0.03	0.01	-0.01	-0.03	-0.01	-0.04
		MAR2	-0.19	-4.69	-0.06	-0.68	-0.04	-0.04	-	-0.04	-0.06	-1.15	-0.05	-0.25	0.02	-1.09	0.05	-0.14
	1	MCAR	0.04	0.04	0.03	0.00	0.07	0.04	-	-2.00	0.02	0.04	0.84	0.50	-0.06	-0.06	0.80	0.41
		MNAR	0.76	0.05	0.05	-0.02	0.05	-0.01	-	-2.32	0.11	0.01	0.84	0.52	0.02	-0.03	0.75	0.44
		MAR	0.01	-0.02	0.01	-0.01	0.07	0.04	-	-2.24	0.06	0.04	0.88	0.56	-0.07	-0.03	0.77	0.44
		MNAR2	0.82	0.02	0.03	0.05	0.04	-0.03	-	-3.51	0.38	0.08	0.89	0.68	0.31	-0.03	0.75	0.57
		MAR2	-0.08	-4.64	-0.01	-0.72	0.03	-1.05	-	-6.19	-0.05	-1.05	0.82	0.47	-0.01	-1.06	0.86	0.48
0.8	0	MCAR	0.00	0.00	0.00	-0.02	-0.01	-0.01	-	0.03	0.00	-0.01	-0.01	-0.03	0.02	0.02	0.01	0.00
		MNAR	-0.01	0.00	0.00	0.02	-0.02	-0.02	-	0.01	-0.01	-0.01	-0.03	-0.04	0.02	0.02	0.01	0.00
		MAR	0.03	0.04	0.03	0.01	-0.02	-0.02	-	0.02	-0.01	-0.03	-0.03	-0.05	0.01	0.04	0.00	0.00
		MNAR2	0.01	0.02	0.03	-0.07	-0.06	-0.06	-	-0.03	0.01	0.01	-0.01	-0.02	0.08	0.08	0.05	0.05
		MAR2	-0.17	-4.46	-0.04	-2.81	-0.04	-0.04	-	-0.02	0.03	-0.61	0.02	-0.41	0.01	-0.63	0.01	-0.43
	1	MCAR	0.00	0.00	0.00	-0.04	0.05	0.03	-	-0.45	-0.04	-0.01	0.95	0.17	0.04	0.04	1.07	0.25
		MNAR	0.29	0.01	0.00	0.03	0.06	0.04	-	-0.47	0.03	0.01	1.04	0.23	-0.01	-0.04	0.98	0.18
		MAR	0.03	0.00	0.03	0.04	0.03	0.00	-	-0.57	0.03	0.04	1.02	0.28	-0.05	-0.04	0.97	0.19
		MNAR2	0.31	0.03	-0.02	0.01	0.03	0.00	-	-0.65	0.00	-0.04	1.02	0.25	0.10	0.05	1.11	0.33
		MAR2	-0.13	-4.29	-0.02	-2.64	-0.04	-0.58	-	-1.34	-0.04	-0.55	0.96	-0.05	0.03	-0.50	1.07	0.02

Table 3.4: Estimation of Liver Cancer Data

	IL*			CC			DV			PB		
	Est.	S.E.	p value	Est.	S.E.	p value	Est.	S.E.	p value	Est.	S.E.	p value
Intercept	2.606	0.378	<.0001	2.601	0.401	<.0001	2.545	0.369	<.0001	2.503	0.370	<.0001
BMI	-0.016	0.633	0.9800	-0.158	0.658	0.8107	-0.0004	0.632	0.9995	0.038	0.630	0.9520
Age	-0.783	0.293	0.0082	-0.706	0.314	0.0260	-0.788	0.292	0.0077	-0.762	0.292	0.0099
Jaundice	0.255	0.123	0.0388	0.236	0.131	0.0744	0.256	0.122	0.0377	0.259	0.123	0.0363
$\sqrt{TSD}$	-0.394	0.512	0.4429	0.013	0.558	0.9816	0	NA	NA	0.001	0.562	0.9940

* IL: ignorable maximum likelihood; CC: complete-case analysis; DV: dropping variable *TSD* ; PB: pseudo-Bayesian analysis.

CHAPTER 4

To model or not to model the missing data mechanism in regression with missing covariates

Abstract: We consider regression with missing covariates. Common methods include: (1) Complete-case analysis (CC), which discards the incomplete cases; (2) Ignorable likelihood methods (IL), which base inference on the observed likelihood given a model for the variables, without modeling the missing data mechanism; (3) Nonignorable modeling (NIM), which bases inference on the joint distribution of variables and the missing data indicators. CC and IL methods do not model the missing data mechanism while NIM models the joint distribution of variables and the missing data indicators. In this paper, we study the question of when it is necessary to model the missing data mechanism. We will study two aspects of covariate missingness on the estimation of regression: (1) nonignorability, which concerns mainly how IL methods perform under varying levels of association between missingness and the missing covariates; (2) outcome dependency, which studies the relatedness of covariate missingness to the outcome on the estimation of regression. We compare different methods for regression with missing covariates using a series of simulation experiments.

Some key words: Complete-case analysis, Ignorable likelihood, nonignorable modeling, outcome dependency

4.1 Introduction

We consider multivariate regression with missing covariates, with data displayed in Figure 4.1. There is a set of outcomes *Y* and two sets of regressor variables *Z* and *W*, with *Z* and *Y* fully observed and *W* with missing values. Here we assume *W* is a single variable, though generalization to multivariate *W* is possible. We denote by (z_i, w_i, y_i) the values of (*Z*, *W*, *Y*) for observation *i*, and by R_{w_i} the indicator for whether *W* is observed or missing. Among of the many reviews of regression with missing covariates are Little (1993), Ibrahim et al. (1999), Ibrahim et al. (2005) and Chen et al. (2008). Common methods include: (1) Complete-case analysis (CC), which discards the incomplete cases; (2) Ignorable likelihood methods (IL), which base inference on the observed likelihood given a model for the distribution of *Y* and *W* given *Z*, without modeling the missing data mechanism; (3) Nonignorable modeling (NIM), which bases inference on the joint distribution of variables and the missing data indicators.

The central problem of this paper is whether to model the missing data mechanism or not in regression with missing covariates. Among the three methods above, CC and IL methods avoid modeling the missing data mechanism, while NIM specifies the joint distribution of the *Y*, *W* and the missing data indicator R_w.

CC analysis is the default method in most software packages. Much of the statistical literature views CC with disfavor since it discards the incomplete cases. However, CC has the advantage of yielding valid inference when the missingness of covariates does not depend on the outcome. This advantage of CC in regression analysis is usually overlooked.

Ignorable likelihood methods have the advantage of retaining all the data, but assume that missing data are missing at random (MAR), in the sense that missingness does not depend on missing values (Rubin 1976, Little and Rubin 2002) which in our setting means the missingness of covariates does not depend on the underlying missing values of the covariates. IL methods are fully efficient for well-specified models and they are also easy to implement since software packages are widely available (e.g., IVEWARE, PROC MI in SAS). Simulation studies show that IL methods are quite robust in the sense that it performs reasonably well even when the MAR assumption is slightly violated (Little and Zhang, 2011). This is because the efficiency gain by using more cases can compensate for the bias resulting from incorrectly ignoring the missing data mechanism.

When the missingness of W is thought to depend on W, IL methods yield biased estimation. Nonignorable modeling methods, which jointly model the distribution of Y, W and R_w, have been proposed (Lipsitz et al. 1999, Huang et al. 2005). There are several disadvantages with nonignorable modeling: (1) the model is not easy to specify correctly and sensitive to model misspecification; (2) the model is generally weakly identified without restrictions on the parameters; (3) there are limited software programs available for nonignorable modeling.

There exist methods for nonignorable missing covariates in regression that do not model the missing data mechanism. Little and Zhang (2011a) propose subsample ignorable likelihood methods (SSIL), which apply IL methods to a subsample and yield valid inference of the regression, for an assumed class of missing data mechanisms. Zhang and Little (2011b) propose a pseudo Bayesian shrinkage approach for regression,

which results in efficient estimation of certain regression coefficients of interest. Both of these methods entail some loss of information.

In this paper, we study the effect of covariate missingness on the estimation of the regression and consider when it is necessary to model the missing data mechanism. We will study two aspects of covariate missingness on the estimation of regression: (1) nonignorability, which concerns mainly on how IL methods perform under varying levels of nonignorability; (2) outcome dependency, which studies the relatedness of covariate missingness to the outcome on the estimation of regression. To jointly model the distribution of (y_i, w_i, R_{w_i}), we use a Bayesian probit selection model, which will be described in section 4.3. In section 4.4, we evaluate both nonignorability and outcome dependency of covariate missingness on the estimation of regression by a series of simulated experiments. In section 4.5, we apply different methods to a liver cancer study. We show by example in section 4.6 that the subsample ignorable likelihood method is actually fully efficient for some special cases.

4.2 The effect of covariate missingness on regression

In this section, we consider a special case of Figure 4.1, where both Y and W are univariate. Suppose the missingness of W depends on

$$W^* = aY + bW + c^T Z + d \tag{4.1}$$

where a, b and d are known scalars and c is a known vector of the same length as z_i.

Nonignorability: b in Eqs. (4.1) can be viewed as a coefficient of nonignorability. When $b = 0$, ignorable likelihood methods are fully efficient. As b moves away from 0,

the ignorability assumption is violated. IL methods performs reasonably well under slight deviation from ignorability but yields poor estimate of the regression if b is too far away from 0.

Outcome dependency: a in Eqs. (4.1) measures how the missingness of W depends on the outcome Y. Outcome dependency is important in regression with missing covariates, since complete-case analysis gives consistent estimate of the regression if $a = 0$. This fact has been explored in Chapter 2 for developing the subsample ignorable likelihood method. In general, there is some loss of information, but we provide an example in section 4.6 in which the subsample ignorable likelihood method is the same as maximum likelihood and thus is fully efficient.

4.3 A Bayesian selection model for regression with missing covariates

We consider the same data structure as in Figure 4.1, where Y and W are univariate. The selection model factorizes the joint distribution of (y_i, w_i, R_{w_i}) as

$$f\ \ y_i, w_i, R_{w_i} \mid z_i; \theta, \psi\ = f\ \ y_i \mid w_i, z_i; \theta\ \ f\ \ w_i \mid z_i; \theta\ \ f\ \ R_{w_i} \mid y_i, w_i, z_i; \psi \qquad (4.2)$$

We model $f\ \ R_{w_i} \mid y_i, w_i, z_i; \psi$ using the probit selection model

$$probit\ \ R_{w_i} = 0 \mid y_i, w_i, z_i; \psi\ = aY + bW + c^T Z + d\ . \qquad (4.3)$$

It is well known that the model is over identified and needs restrictions to be estimable, in a frequentist setting. One possible restriction is to set the coefficient a to be 0. In a Bayesian setting, all parameters can be estimated without restriction, but the estimation might be very poor in the sense that the MCMC chain has a convergence

problem and the parameters are estimated with large variance (Preget and Waelbroeck, 2006, Freedman and Sekhon, 2010). In this paper, we use a Bayesian model with restriction that a is always 0, and use noninformative prior for all parameters.

Beside the identifiability issue in eq. (4.3), there is also potential instability in estimating the parameters (Little 1985). This results in very poor convergence of the Markov chains. We use the accelerated Gibbs sampler in Omori (2007) to speed up the convergence.

4.4 Simulation

In this section, we evaluate the effect of nonignorability and outcome dependency of covariate missingness on the regression with missing covariates.

We simulate

$$(w_i, z_{1i}, z_{2i}) \sim_{\text{ind}} N(0, \Sigma),$$

where $N(\mu, \Sigma)$ denotes the normal distribution with mean μ and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix},$$

for $i = 1,\ldots,100$. Y is related to w_i, z_{1i}, z_{2i} by the linear model

$$y_i \mid w_i, z_{1i}, z_{2i} \sim_{\text{ind}} N(1 + w_i + z_{1i} + z_{2i}, 1).$$

Missing values of W were then generated based on the following probit model:

$$\text{probit } P(R_{w_i} = 0 \mid w_i, z_{1i}, z_{2i}, y_i) = 1 + bw_i + ay_i + z_{1i} + z_{2i}$$

with z_i fully observed when w_i is missing.

We vary a and b to assess the effect of nonignorability and outcome dependency of covariate missingness and look at the following two sets of simulation:

(I). Ignorable and varying outcome dependency: $b = 0, a = 0, 0.25, 0.5, 1, 2, 4, 8$.

(II). No outcome dependency and varying nonignorability: $a = 0, b = 0, 0.25, 0.5, 1, 2, 4, 8$.

We simulate data for correlation coefficient $\rho = 0$ and $\rho = 0.7$. Four methods are applied to estimate the regression coefficients $\beta_0, \beta_w, \beta_{z1}, \beta_{z2}$:

(1) BD: estimates from the regression before deletion (BD), as a benchmark method.

(2) CC: Complete-case analysis;

(3) IL: ignorable maximum likelihood method assuming MAR;

(4) NIM: nonignorable modeling described in section 4.3.

We report the RMSEs, confidence coverage and empirical bias of the estimated regression coefficients from each method. Results are based on 1000 repetitions for each combination of a and b.

CC analysis gives valid estimate of the regression if the missingness does not depend on the outcome, i.e., $a = 0$. IL gives valid and efficient estimate of the regression if the missing data mechanism is MAR, i.e., $b = 0$.

As we can see from Figure 4.3, 4.4, and 4.5, CC analysis breaks down quickly as a moves away from 0, leading to biased estimate of all regression coefficients. IL is valid and gives the smallest RMSEs among CC, IL, and NIM. The NIM method yields good estimate for the intercept and the regression coefficient of *Z*, but biased estimate for the regression coefficient of *W*, since NIM model restricts the coefficient of y in the probit model to be zero and therefore is incorrectly specified.

In the second set of simulations, CC is a valid method since the missingness does not depend on the outcome. The IL method yields poor estimate of the regression as b moves away from 0. When b is less than 1, IL performs reasonably well for all regression coefficients.

In both sets of simulations, we see an advantage of using the NIM method over the CC method. In the first set of simulations when CC is biased, the NIM is clearly better. CC analysis is unbiased in the second set of simulations; however, there is big efficiency gain of the NIM method over CC analysis because it uses the full sample.

4.5 Application: A liver cancer study

We apply the nonignorable modeling method to the liver cancer dataset in CHAPTER 3. The dataset contains 191 patients from Eastern Cooperative Oncology Group clinical trials EST 2282 (Falkson et al., 1990) and EST 1286 (Falkson et al., 1995). We are interested in how the number of the cancerous liver nodes (CNTs) is predicted by four baseline characteristics:

(1) body mass index (BMI, in kg/m^2);

(2) age (in years);

(3) jaundice (yes, no): the yellowish staining of the skin and the whites of the eye;

(4) time since diagnosis of the disease (TSD, in weeks).

Like many other empirical studies, this dataset contains missing values. TSD is missing for 17 patients (8.9%) while other variables are fully observed. CC analysis suffers from inefficiency and potential bias if the missingness of TSD depends on the outcome. IL makes use of the partial information in the incomplete case but assumes the

missing data are missing at random (MAR; Rubin 1976b, Little and Rubin 2002). NIM jointly models the variables and missing data mechanism, but restricts the coefficient of the outcome CNTs in the selection model to zero, for identifiability purpose. We also apply another version of NIM, which does not restrict the coefficient of the outcome to be zero in modeling the missing data mechanism (NIM-Y). NIM-Y is not identified in a frequestist setting, but parameters can be estimated using posterior simulation in a Bayesian setting. NIM-Y has the virtue of correctly specifying the model if missingness does depend on the outcome.

For the liver cancer example, NIM indicates that longer TSD is associated with missingness though the estimate is not significant. NIM-Y shows that missingness can be predicted by the outcome CNTs (with an estimate of 0.44 and 95% C.I. (0.10, 0.83)), implying that CC analysis might lead to biased estimate. The coefficients of TSD in the selection part of NIM and NIM-Y are estimated with large variance because the information about the unobserved TSD is scarce.

Table 4.1 shows the regression coefficients as well as the 95% C.I.s (confidence intervals or credible intervals). CC has a larger regression coefficient estimate of Age and smaller estimate of Jaundice. This is not surprising since NIM-Y shows that missingness depends on the outcome, and therefore CC analysis leads to biased estimate of the regression. IL gives a smaller estimate of TSD compared to the other three methods, which might be explained by the positive (though not significant) association between missingness and TSD.

4.6 A Normal Regression Model where SSIML is ML

The subsample ignorable likelihood method in CHAPTER 2 can be viewed as complete-case analysis on a certain set of variables. Generally, there is some loss of information. However in some special cases, the proposed method is full maximum likelihood and hence fully efficient. We give an example, an extension of Example 2 in Little and Wang (1996).

Consider the special case of Figure 2.3 shown in Figure 4.2, where *W*, *X* and *Y* (but not necessarily *Z*) are univariate, Z and Y are fully observed, *X* is missing and *W* is observed in Pattern 2, and *W* is missing and *X* is observed in Pattern 3. Restating assumptions (2.11) and (2.12) in this special case yields:

$$p\ \ R_{w_i}=1|\, z_i, w_i, x_i, y_i, \psi_w\ \ = p\ \ R_{w_i}=1|\, z_i, w_i, x_i, \psi_w\ \ \text{for all } y_i \tag{4.4}$$

$$p(R_{x_i}=1|\, z_i, w_i, x_i, y_i, R_{w_i}=1, \psi_{xy\cdot w}) = p(R_{x_i}=1|\, z_i, w_i, y_i, R_{w_i}=1, \psi_{xy\cdot w}) \text{ for all } x_i \tag{4.5}$$

We model the joint distribution of *W*, *X*, *Y*, R_X and R_W given *Z* as follows:

$$p(w_i, x_i, y_i, R_{w_i}, R_{x_i} \mid z_i, \theta, \psi_w, \psi_{x\cdot w}) =$$
$$p(w_i, x_i, y_i \mid R_{w_i} = j, z_i, \theta^{(j)}) \times p(R_{w_i} \mid z_i, \psi_w) \times p(R_{x_i} \mid R_{w_i}, y_i, w_i, x_i, z_i, \psi_{x\cdot w}),$$

where the three sets of parameters $(\theta, \psi_w, \psi_{x\cdot w})$ are distinct, $(w_i, x_i, y_i \mid z_i, R_{w_i} = j, \theta^{(j)})$ are assumed to have a trivariate normal distribution with mean $\beta_0^{(j)} + \beta_z^{(j)} z_i$ and covariance matrix $\Sigma^{(j)}$, and $\theta = (\theta^{(0)}, \theta^{(1)})$ where $\theta^{(j)} = \beta_0^{(j)}, \beta_z^{(j)}, \Sigma^{(j)}$, $j = 0,1$. The models for R_{w_i}, R_{x_i} are left arbitrary, subject to the distinctness of parameters. The observed likelihood for this model is

$$L_{\text{obs}}(\theta,\psi_w,\psi_{x\cdot w}) = \prod_{R_{w_i}=1} P\ \ R_{w_i}=1 \mid z_i,\psi_w\ \times \prod_{R_{w_i}=0} P\ \ R_{w_i}=0 \mid z_i,\psi_w\ \times$$

$$\prod_{R_{w_i}=R_{x_i}=1} P\ \ w_i,x_i,y_i \mid z_i,R_{w_i}=1,\theta^{(1)}\ \times \prod_{R_{w_i}=1,R_{x_i}=0} P\ \ w_i,y_i \mid z_i,R_{w_i}=1,\theta^{(1)}\ \times$$

$$\prod_{R_{w_i}=1,R_{x_i}=0} \left[\int P\ \ w_i,x,y_i \mid z_i,R_{w_i}=1,\theta^{(1)}\ \ P\ \ R_{x_i}=0 \mid y_i,w_i,z_i,R_{w_i}=1,\psi_{x\cdot w}\ \ dx\right]\times$$

$$\prod_{R_{w_i}=0,R_{x_i}=1} P\left(x_i,y_i \mid z_i,R_{w_i}=0,\theta^{(0)}\right)$$

Under the subsample MAR condition (4.5), the third line factorizes, yielding

$$L_{\text{obs}}(\theta,\psi_w,\psi_{x\cdot w}) = L_1(\psi_w)\times L_2(\theta^{(1)})\times L_3(\theta^{(0)})\times L_4(\psi_{x\cdot w}),\ \text{where:}$$

$$L_1(\psi_w) = \prod_{R_{w_i}=1} P\ \ R_{w_i}=1 \mid z_i,\psi_w\ \times \prod_{R_{w_i}=0} P\ \ R_{w_i}=0 \mid z_i,\psi_w\ ,$$

$$L_2(\theta^{(1)}) = \prod_{R_{w_i}=R_{x_i}=1} P\ \ w_i,x_i,y_i \mid z_i,R_{w_i}=1,\theta^{(1)}\ \times \prod_{R_{w_i}=1,R_{x_i}=0} P\ \ w_i,y_i \mid z_i,R_{w_i}=1,\theta^{(1)}$$

$$\times \prod_{R_{w_i}=1,R_{x_i}=0} P\ \ w_i,x,y_i \mid z_i,R_{w_i}=1,\theta^{(1)}\ ,$$

$$L_3(\theta^{(0)}) = \prod_{R_{w_i}=0,R_{x_i}=1} P\left(x_i,y_i \mid z_i,R_{w_i}=0,\theta^{(0)}\right),$$

$$L_4(\psi_{x\cdot w}) = \prod_{R_{w_i}=1,R_{x_i}=0} P\left(R_{x_i}=0 \mid y_i,w_i,z_i,R_{w_i}=1,\psi_{x\cdot w}\right).$$

Subsample ignorable ML (SSIML) maximizes, L_2, yielding ML estimates of $\theta^{(1)}$, the parameters of the distribution of W,X,Y given Z for cases with W observed. Write $\theta^{(0)} = (\theta^{(0)}_{XY\cdot Z},\theta^{(0)}_{W\cdot XYZ})$, where $\theta^{(0)}_{XY\cdot Z}$ are the parameters of the distribution of X,Y given Z and $\theta^{(0)}_{W\cdot XYZ}$ are the parameters of the regression of W on X,Y,Z, both for cases with W missing. Maximizing L_3 yields ML estimates of $\theta^{(0)}_{XY\cdot Z}$, but the remaining components $\theta^{(0)}_{W\cdot XYZ}$ do not appear in the likelihood. However, they are just identified by assumption (4.4), which implies that $\theta^{(0)}_{Y\cdot WXZ} = \theta^{(1)}_{Y\cdot WXZ} = \phi$, the parameters of the regression of interest.

This identification by parameter restrictions extends the analysis of Little and Wang (1996) to this more complex pattern.

It follows that the SSIML estimate of ϕ obtained from estimating $\theta^{(1)} = \hat{\theta}^{(1)}$ to maximize L_2, and setting $\hat{\phi} = \phi(\hat{\theta}^{(1)})$, is ML for this model, subject to one caveat: the resulting ML estimates of $\theta^{(0)}$ must lie in the parameter space, and in particular the covariance matrix must be positive definite, a condition that is not guaranteed by the transformation (Little and Wang 1996). If they do not, SSIML is still consistent but not the same as ML. Bayesian inference based on L_2 with a prior distribution for the parameters is not the same as fully Bayesian inference, since draws of the posterior distribution of $\theta^{(0)}$ need to be restricted to lie in their parameter space.

The key to this argument is the fact that the number of unidentified parameters in $\theta^{(1)}_{W \cdot XYZ}$ equals the number of restrictions in $\theta^{(0)}_{Y \cdot WXZ} = \theta^{(1)}_{Y \cdot WXZ} = \phi$, yielding a (1-1) transformation between the two parameter sets. This generalizes to cases where *W* and *Y* are normal with the same dimension, but not to cases where *W* and *Y* have different dimension.

4.7 Conclusion

This paper looks at two aspects of covariate missingness in regression analysis: (1) nonignorability, which concerns whether the missingness depends on the underlying missing values; (2) outcome dependency, which relates the missingness to the outcome. We use a series of simulation to study the effect of nonignorability and outcome dependency on CC analysis and IL analysis for handling missing covariates in regression

analysis. When the missingness is ignorable, IL method yields efficient estimate of the regression. IL method is robust to slight violation of ignorability, in the sense that it gives estimate with small RMSEs compared to CC analysis. Simulation shows that CC analysis breaks down rapidly if the covariate missingness depends on the outcome.

In this paper, we show the results of different methods for multiple linear regression with one incomplete covariate. The analysis can be generalized to multiple linear regression with more covariates missing; also similar analysis can be done for generalized linear models and survival analysis.

It is important to note that neither of nonignorability and outcome dependency are testable from the data, and therefore the nonignorable modeling (NIM) considered in this paper cannot be viewed as a test of nonignorability or outcome dependency. Different forms of nonignorable models can be used as a sensitivity analysis in real data analysis, for example, pattern mixture model (Little and Wang 1996) and the two selection models described in this paper.

Figure 4.1 Missing Data Structure in Section 4.1

Pattern	Observation, i	z_i	w_i	y_i	R_{w_i}
1	$i = 1,\ldots,m$	√	√	√	1
2	$i = m+1,\ldots,n$	√	x	√	0

Key: √ denotes observed, x denotes missing

Figure 4.2 Missing Data Structure for Section 4.5

Pattern	Observation, i	z_i	w_i	x_i	y_i	R_{w_i}	R_{x_i}
1	$i = 1,\ldots,m$	√	√	√	√	1	1
2	$i = m+1,\ldots,m+r$	√	√	x	√	1	0
3	$I = m+r+1,\ldots,n$	√	x	√	√	0	1

Key: √ denotes observed, x denotes missing.

Figure 4.3 : RMSE: Ignorable – outcome dependency varies (ρ=0)

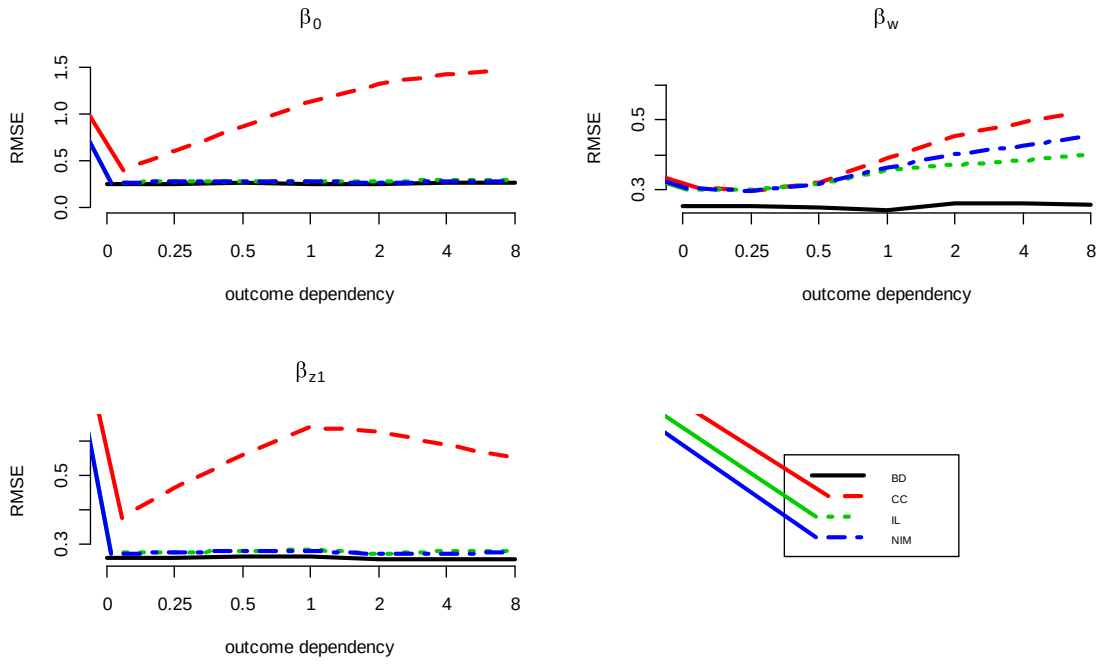

Figure 4.4 : Coverage: Ignorable – outcome dependency varies(ρ=0)

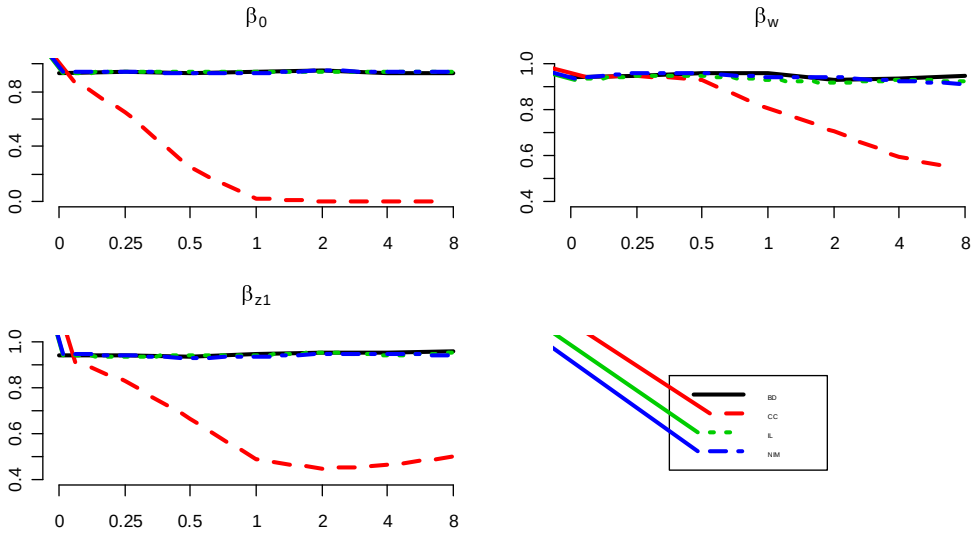

Figure 4.5 : Bias: Ignorable – outcome dependency varies(ρ=0)

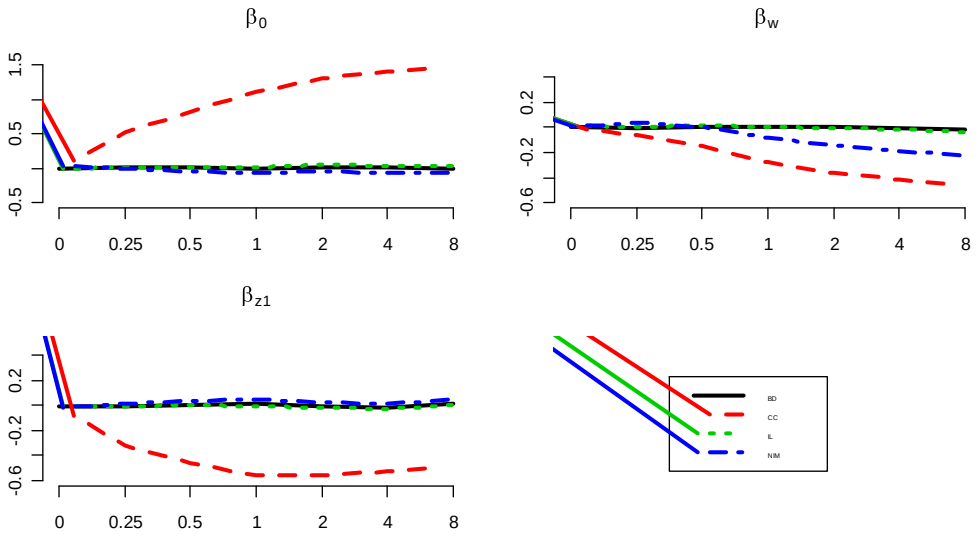

Figure 4.6 : RMSE: No outcome dependency – Nonignorability varies(ρ=0)

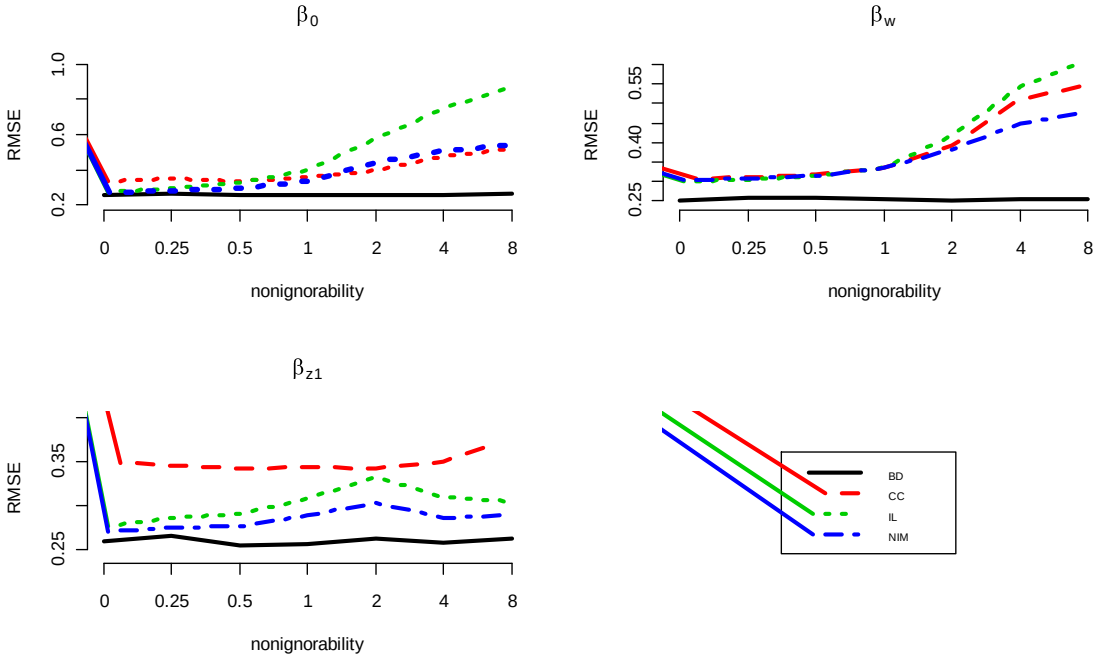

Figure 4.7 : Coverage: No outcome dependency – Nonignorability varies(ρ=0)

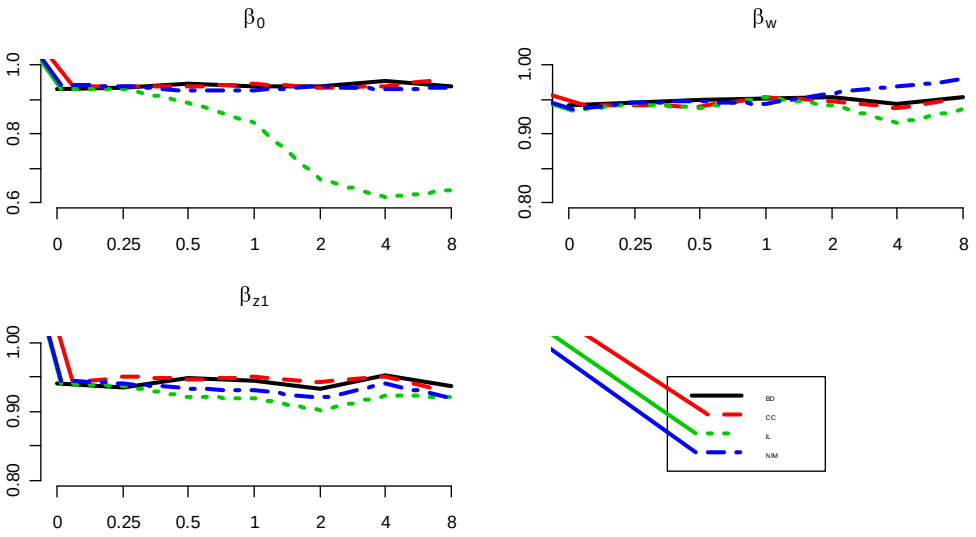

Figure 4.8 : Bias: No outcome dependency – Nonignorability varies (ρ=0)

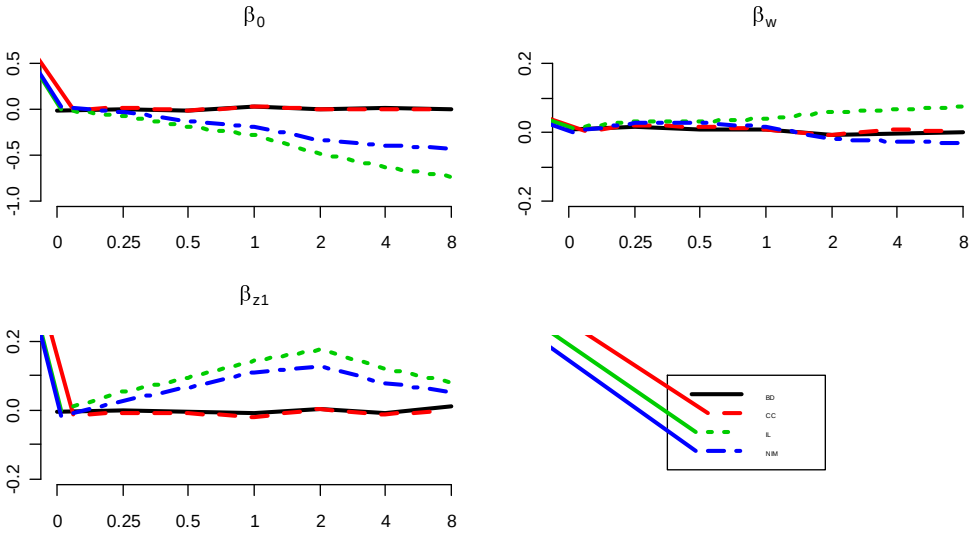

Figure 4.9: RMSE: Ignorable-outcome dependency varies (ρ=0.7)

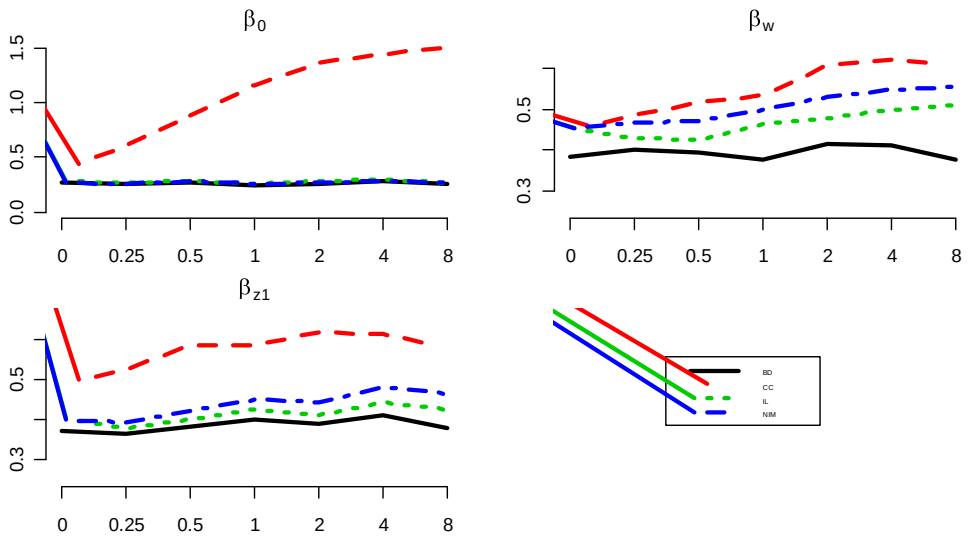

Figure 4.10: Coverage: Ignorable – outcome dependency varies(ρ=0.7)

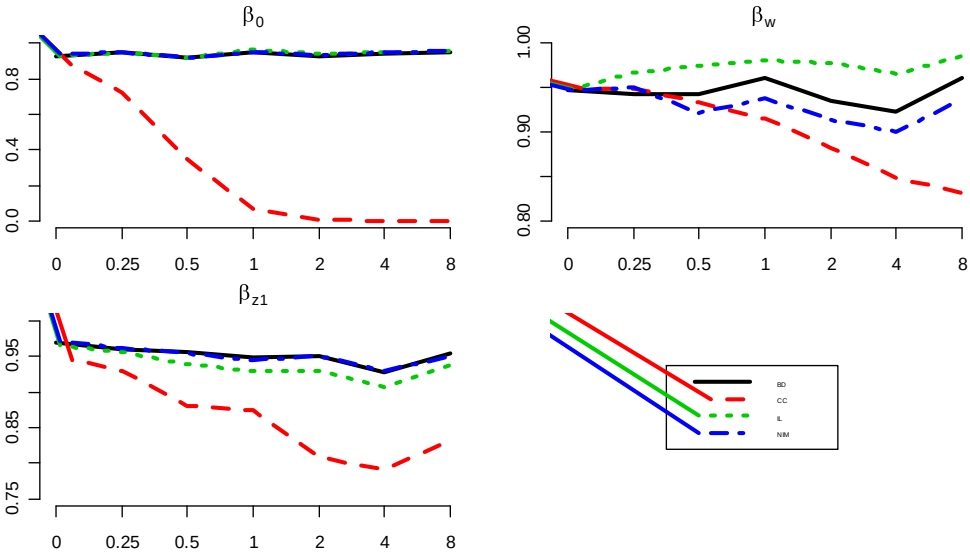

Figure 4.11: Bias: Ignorable – outcome dependency varies(ρ=0.7)

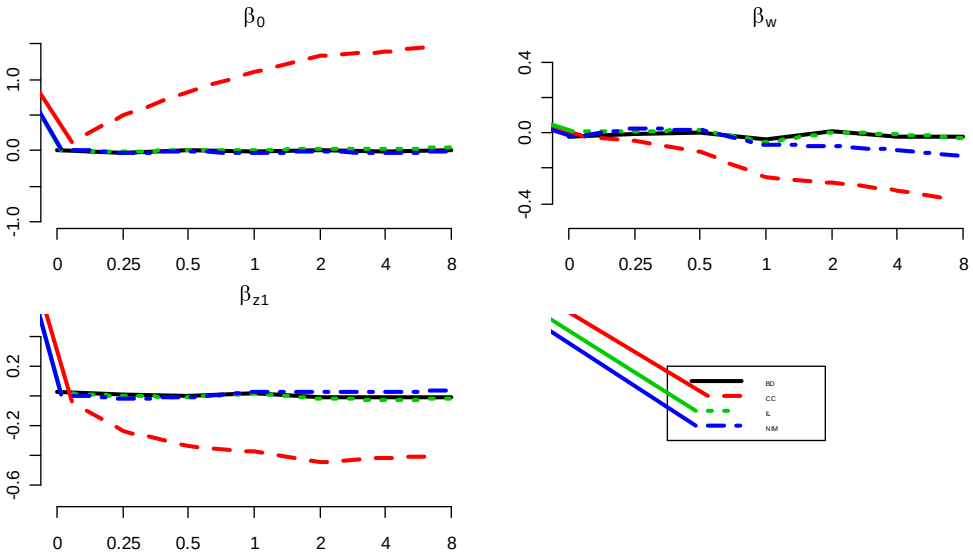

Figure 4.12: RMSE: No outcome dependency – Nonignorability varies(ρ=0.7)

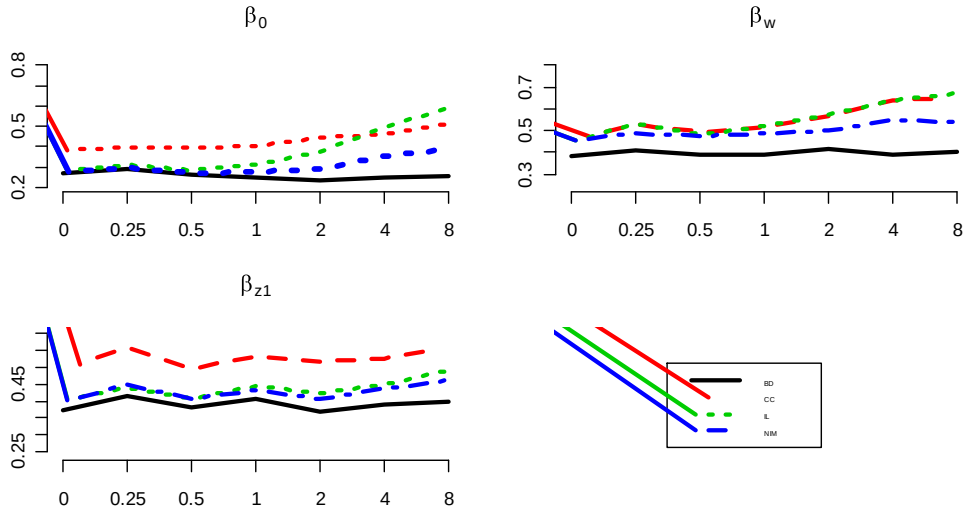

Figure 4.13: Coverage: No outcome dependency – Nonignorability varies(ρ=0.7)

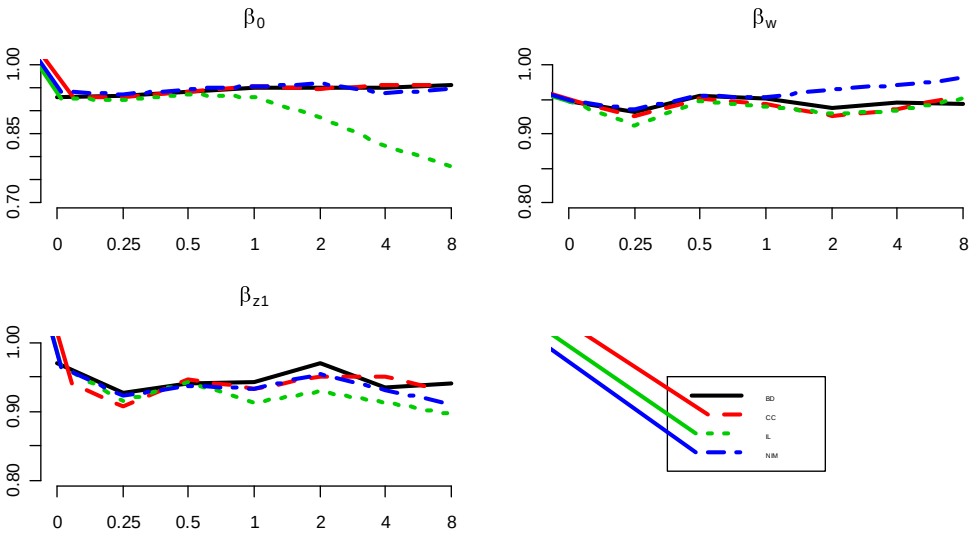

Figure 4.14: Bias: No outcome dependency – Nonignorability varies (ρ=0.7)

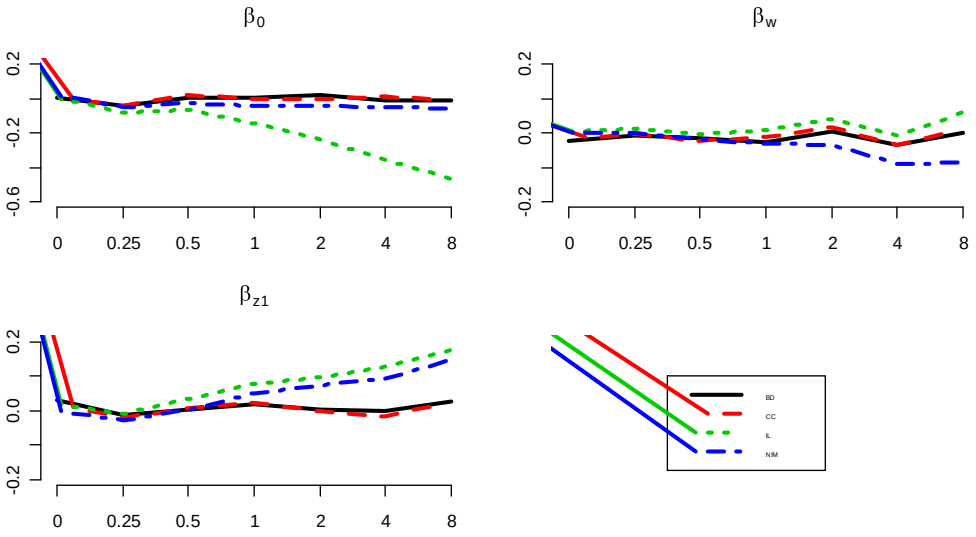

Table 4.1: Estimation of Liver Cancer Data

	IL*		CC		NIM^		NIM-Y‡	
	Est.	95% C.I.	Est.	95% C.I.	Est.	95% C.I.	Est.	95% C.I.
Intercept	2.606	(1.865, 3.347)	2.601	(1.815, 3.387)	2.538	(1.750, 3.317)	2.544	(1.753, 3.328)
BMI	-.016	(-1.257, 1.225)	-.158	(-1.448, 1.132)	-.004	(-1.408, 1.331)	-.011	(-1.328, 1.377)
Age	-.783	(-1.357, -.209)	-.706	(-1.321, -.091)	-.784	(-1.398, -0.160)	-.785	(-1.385, -.186)
Jaundice	.255	(.014, .496)	.236	(-.021, .493)	.255	(-.006, .506)	.252	(-.006, .508)
$\sqrt{TSD}$	-.394	(-1.398, .610)	.013	(-1.081, 1.107)	.046	(-1.085, 1.146)	.055	(-1.063, 1.210)

* IL: ignorable maximum likelihood; CC: complete-case analysis; NIM: dropping nonignorable modeling with restrition; NIM-Y: nonignorable modeling with no restriction.

^: Coefficient and 95% C.I. in the NIM selection model: Intercept: 1.22 (-0.54, 3.05); BMI: 0.40 (-2.19, 3.34); Age: -0.52 (-1.75, 0.69) Jaundice: 0.28 (-0.21, 0.77); TSD: 1.97 (-4.65, 8.31).

‡: Coefficient and 95% C.I. in the NIM-Y selection model: Intercept: 0.38 (-1.54, 2.34); BMI: 0.14 (-2.50, 3.10); Age: -0.11 (-1.44, 1.15); Jaundice: 0.18 (-0.35, 0.70); TSD: 0.97 (-5.78, 7.46); CNTs: 0.44 (0.10, 0.83).

CHAPTER 5

| Conclusions and Future Work

We consider regression with missing covariates in this dissertation. When the missing data mechanism is missing at random, the ignorable likelihood method is the most efficient method. When the missing data mechanism is missing not at random, IL methods are biased. One possibility is to apply a nonignorable modeling method, but such methods are vulnerable to misspecification of the missing data mechanism, and suffer from problems with identifying the parameters. In Chapter 2 and Chapter 3, we propose two methods that do not model the missing data mechanism, the subsample ignorable likelihood method (SSIL) and the pseudo-Bayesian Shrinkage method (PB), both of which yield estimate with nice properties under certain circumstances. In Chapter 4, we use a series of simulated experiments to evaluate the effect of nonignorability and outcome dependency of covariate missingness on two common methods: the complete-case analysis (CC) and the IL method.

In Chapter 2, we propose the subsample ignorable likelihood (SSIL) method, which applies an IL method to the subsample of observations that are complete on one set of variables, but possibly incomplete on others. We give the conditions on the missing data mechanism under which SSIL gives consistent estimates, but both complete-case analysis and IL methods are inconsistent. The general theoretical rationale of SSIL is partial likelihood (Cox, 1972). This involves a potential loss of efficiency relative to full modeling, but we show in Chapter 4 an example in which SSIL is fully efficient.

We present the SSIL method in a likelihood setting but it is also applies to non-likelihood analyses that are valid under the MAR assumption. For example, for repeated-measures data, the IL method applied to the subsample could be replaced by a method such as weighted generalized estimating equations (WGEE), which is also valid under MAR, without affecting the validity of the method under the stated assumptions (2.7) and (2.8).

It is worthwhile to apply the subsample method to the proportional hazards model (PHReg) with missing covariates. Both the PHReg and the subsample method are partial likelihood, and it is interesting to see how a new method that combines these two works.

We will also apply the subsample ignorable likelihood method to longitudinal surveys, in which for the subsample that are complete in previous surveys, missingness of subsequent survey items may be assumed to depend on the observed data, like the subsample MAR assumption in Chapter 2. The subsample MAR assumption is less stringent than the MAR assumption, and therefore the subsample ignorable likelihood method maybe a preferred method than the multiple imputation (MI) method, which assumes MAR.

In Chapter 3, we propose a pseudo-Bayesian shrinkage method for regression analysis with a missing covariate, which is a compromise between complete-case analysis and the analysis that drops the missing covariate. The method recovers information in the incomplete cases by assigning the regression coefficient of the incomplete variable a mixture prior of a normal distribution and a point mass at zero.

In future work, we will extend the pseudo-Bayesian shrinkage method to more than one missing covariate, and other parametric regression models, like generalized linear models and survival analysis.

The proposed method could be combined with existing multiple imputation methods to handle more general problems where Z is also incomplete. In particular, when missingness of covariates W is MNAR but does not depend on the outcome, and missingness of Z is MAR, the method could also be applied by assigning similar mixture priors to the regression coefficients of W, while using multiple imputation via chained equations (Raghunathan et al., 2001; IVEware, 2011; MICE, 2011) to impute missing values of Z.

In Chapter 4, we study two aspects of covariates missingness, the nonignorability and outcome dependency. We compare different methods under varied levels of nonignorability and outcome dependency using a series of simulated experiments. Simulation shows that CC analysis performs poorly even under slight violation of outcome dependency. IL method is most efficient when the missing data mechanism is ignorable and is also robust to slight violation of ignorability. For future work, it is interesting to extend the analysis to multiple regression with more than one covariate missing and other parametric models with missing covariates.

We generate the missing data based on probit selection models and use the correct model to model the missing data mechanism. It is interesting to see how the selection model performs when the selection model is not specified correctly. For future work, we will look at different violations of an additive probit model, for example, generating the

missing data indicator from a heavy-tailed distribution or including a nonlinear or interaction term in the missing data generation scheme.

REFERENCES

Amemiya, T. (1984). Tobit models, a survey. *J. Econometrics* 24, 3-61.

Centers for Disease Control and Prevention (CDC) (2004). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Chen, Q., Ibrahim, J.G., Chen, M.H., and Senchaudhuri, P. (2008). Theory and Inference for Regression Models with Missing Responses and Covariates. *Journal of Multivariate Analysis* 99, 1302-1331.

Chen, Q., Zeng, D. and Ibrahim, J.G. (2007). Sieve Maximum Likelihood Estimation for Regression Models with Covariates Missing at Random. *J. Am. Statist. Assoc.* 102, 1309-1317.

Chen, Q., Ibrahim, J.G., Chen, M.H., and Senchaudhuri, P. (2008). Theory and Inference for Regression Models with Missing Responses and Covariates. *Journal of Multivariate Analysis* 99, 1302-1331.

Colhoun, H., Hemingway, H., Poulter, N.R. (1998). Socio-economic status and blood pressure: an overview analysis. *Journal of Human Hypertension* 12, 91-110.

Cox, D. R. and Reid, N. (1987). Parameter Orthogonality and Approximate Conditional Inference. *Journal of the Royal Statistical Society B* 49, 1-30.

Das, U., Maiti, T., and Pradhan, V. (2010). Bias correction in logistic regression with missing categorical covariates. *Journal of Statistical Planning and Inference* 140, 2478-2485.

David, M., Little, R. J. A., Samuhel, M.E. and Triest, R. K. (1986) Albernative Methods for CPS Income Imputation. *J. Am. Statist. Assoc.* 86, 29-41.

Falkson, G., Cnaan, A., and Simson, I.W. (1990). A randomized phase II study of activicim and 4'deoxydoxorubicinain patients with hepatocellular carcinoma in an Eastern Cooperative Oncology Group Study. *American Journal of Clinical Oncology* 13, 510-515.

Falkson, G., Lipsitz, S., Borden, E., Simson, I.W., and Haller, D. (1995). A ECOG randomized phase II study of beta interferon and Menogoril. *American Journal of Clinical Oncology* 18, 287-292.

Freedman, D.A. and Sekhon, J.S. (2010). Endogeneity in Probit Response Models. *Political Analysis* 10, 138-150.

George, E.I. (2000). The Variable Selection Problem. *J. Am. Statist. Assoc.* 95, 1304-1308.

George, E.I. and McCulloch, R.E. (1993). Variable Selection Via Gibbs Sampling. *J. Am. Statist. Assoc.* 88, 881-889.

George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian Variable Selection. *Statistics Sinica* 7, 339-373.

Gilks, W.R., Thomas, A., and Spiegelhalter, D.J. (1994). A language and program for complex Bayesian modeling. *The Statistician* 43, 169-178.

Glynn, R. J., and Laird, N. M. (1986). Regression Estimates and Missing Data: Complete-Case Analysis. Technical Report, Harvard School of Public Health, Dept. of Biostatistics.

Goffinet, B. (1987). Alternative conditions for ignoring the process that causes missing data. *Biometrika* 74, 437-439.

Gulliford, M.C., Mahabir, D. and Rocke, B. (2004). Socioeconomic inequality in blood pressure and its determinants: cross-sectional data from Trinidad and Tobago. *Journal of Human Hypertension* 18, 61-70.

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. *Ann. Econ. Soc. Meas.*. 5, 475-492.

Huang, L., Chen, M.H., and Ibrahim, J.G. (2005). Bayesian Analysis for Generalized Linear Models with Nonignorably Missing Covariates. *Biometrics* 61, 767-780.

Ibrahim, J.G, Chen, M.H., and Lipsitz, S.R. (1999). Monte Carlo EM for Missing Covariates in Parametric Regression Models. *Biometircs* 55, 591-596.

Ibrahim, J.G, Chen, M.H., and Lipsitz, S.R. (2002). Bayesian Methods for Generalized Linear Models with Covariates Missing at Random. *Canadian Journal of Statistics* 30, 55-78.

Ibrahim, J.G., Chen, M. H., Lipsitz, S.R, and Herring, A.H. (2005). Missing Data Methods for Generalized Linear Models: A Comparative Review. *J. Am. Statist. Assoc.* 100, 332-346.

Ibrahim, J.G., Lipsitz, S.R., and Chen, M.H. (1999). Missing Covariates in Generalized Linear Models When the Missing Data Mechanism is Nonignorable. *Journal of the Royal Statistical Society B* 61, 173-190.

Jones, M.P. (1996). Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. . *J. Am. Statist. Assoc.* 91, 222-230.

Kuo, L. and Mallick, B. (1998). Variable Selection for Regression Models. *Sankhya Series B* 60, 65-81.

Kim, S., Egerter, S., Cubbin, C., Takahashi, E. R., and Braveman, P. (2007). Potential Implications of Missing Income Data in Population-Based Surveys: An Example from a Postpartum Survey in California. *Public Health Rep.* 112, 753-763.

Lillard, L., Smith, J. P. and Welch, F. (1986). What do We Really Know About Wages: The Importance of Nonreporting and Census Imputation. *Journal of Political Economy* 94, 489-506.

Little, R.J.A. (1979). Maximum likelihood inference for multiple regression with missing values: a simulation study. *Journal of the Royal Statistical Society, Series B*, 41, 76-87.

Little, R.J.A. (1985). A Note about Models for Selectivity Bias. *Econometrica* 53, 1469-1474.

Little, R.J.A. (1992). Regression with Missing X's: A Review. *J. Am. Statist. Assoc.* 87, 1127-1137.

Little, R. J. A (1993). Pattern-Mixture Model for Multivariate Incomplete Data. *J. Am. Statist. Assoc.* 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., and Rubin, D. B. (2002). *Statistical Analysis with Missing Data (2nd ed)*. New Jersey: John Wiley.

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

Little, R.J.A., and Zhang, N. (2010). Subsample Ignorable Likelihood for Regression with Missing Data. Submitted for publication.

Mackenbach, J.P. (1994). The epidemiologic transition theory. *J. Epidemiol Community Health* 48, 329-331.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd ed., London: Chapman and Hall.

Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian Variable Selection in Linear Regression. *J. Am. Statist. Assoc.* 83, 1023-1032.

Omori, Y. (2007). Efficient Gibbs sampler for Bayesian analysis of a sample selection model. *Statistics & Probability Letters* 77, 1300-1311.

Preget, R. and Waelbroeck, P. (2006). Sample Selection with Binary Endogenous Variable: A Bayesian Analysis of Participation to Timber Auctions. *Telecom Paris Economics and Social Sciences Working Paper No. ESS-06-08.*

Raghunathan, T., Lepkowski, J. VanHoewyk, M., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Method.* 27, 85-95. For associated IVEWARE software see http://www.isr.umich.edu/src/smp/ive/.

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

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika* 63, 581-592.

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

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

SAS (2010). Statistical Analysis with SAS/STAT® Software, http://www.sas.com/technologies/analytics/statistics/stat/index.html

Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G., and Cohen, A.J. (2006). Multiple Imputation of Missing Income Data in the National Health Interview Survey. *J. Am. Statist. Assoc.* 101, 924-933.

Tolley, G.S., and Olson, E. (1971). The Interdependence between Income and Education. *Journal of Political Economy* 79, 460-480.

Von Hippel, P. T. (2007). Regression with Missing Ys: an Improved Strategy for Analyzing Multiply Imputed Data. *Sociological Methodology*, 37, 1, 83-117.

Yan, T., Curtin, R. and Jans. M. (2010). Trends in Income Nonresponse Over Two Decades. *Journal of Official Statistics* 26, 145-164.

Zhang, N. and Little, R.J. (2011). A Pseudo Bayesian Shrinkage Approach to Regression with Missing Covariates. Submitted for publication.