Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2011)
60, *Part* 4, *pp.* 591–605

Subsample ignorable likelihood for regression analysis with missing data

Roderick J. Little and Nanhua Zhang

University of Michigan, Ann Arbor, USA

[Received June 2010. Revised December 2010]

Summary. Two common approaches to regression with missing covariates are complete-case analysis and ignorable likelihood methods. We review these approaches and propose a hybrid class, called subsample ignorable likelihood methods, which applies an ignorable likelihood method to the subsample of observations that are complete on one set of variables, but possibly incomplete on others. Conditions on the missing data mechanism are presented under which subsample ignorable likelihood gives consistent estimates, but both complete-case analysis and ignorable likelihood methods are inconsistent. We motivate and apply the method proposed to data from the National Health and Nutrition Examination Survey, and we illustrate properties of the methods by simulation. Extensions to non-likelihood analyses are also mentioned.

Keywords: Maximum likelihood; Missing data; Multiple imputation; Multivariate regression; Non-ignorable data mechanism

1. Introduction

Missing data are 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 (1992), Ibrahim *et al.* (1999, 2002, 2005) and Chen *et al.* (2008). Three approaches are

(a) complete-case (CC) analysis, 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 (Rubin, 1987), as in SAS `PROC MI` (SAS Institute, 2010) or IVEware (Raghunathan *et al.*, 2001)) and
(c) non-ignorable modelling, 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 the 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 they assume that 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;

Address for correspondence: Roderick J. Little, Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109-2029, USA.
E-mail: rlittle@umich.edu

 0035–9254/11/60591

Little and Rubin, 2002). CC analysis involves a loss of information but has the advantage of yielding valid inferences when missingness depends on the missing covariates X but not the response Y, a potentially non-ignorable mechanism where IL methods are subject to bias. This advantage of CC analysis 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 paper shows that the answer is yes, under particular assumptions about the missing data mechanism which are formalized in Section 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 method proposed, subsample ignorable likelihood (SIL), then applies an IL method to the subsample of cases with W observed. Particular forms that are discussed below are subsample ignorable maximum likelihood (SIML), which applies IML to the subsample, and subsample ignorable multiple imputation (SIMI), which applies an ignorable data model to multiply-impute the missing values in the subsample.

Section 2 presents a motivating application based on data from the National Health and Nutrition Examination Survey (NHANES) (Centers for Disease Control and Prevention, 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 that the missingness of education, BMI and the blood pressure measures was at random, 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 3 reviews properties of CC and IL analyses, and Section 4 presents properties of the proposed SIL methods. In particular, conditions on the missing data mechanism are presented under which SIL gives consistent estimates, but both IL and CC analyses are inconsistent. In other circumstances, IL is inconsistent and SIL and CC analysis are consistent, but SIL is more efficient than CC analysis since it uses more of the data. Section 5 presents simulations that illustrate the properties of SIL and alternative methods. In Section 6 we apply the method to the motivating data from the NHANES (Centers for Disease Control and Prevention, 2004). We conclude with some discussion in Section 7.

2. Motivating problem

The effect of socio-economic status on blood pressure has been studied by many researchers (Gulliford *et al.* (2004); Colhoun *et al.* (1998), etc.). The results provide an important basis for public health interventions. The effect of socio-economic status on blood pressure generally varies by geographical region and time as the risk factors in populations change (Mackenbach, 1994). The data set that is analysed in this paper is from the 2003–2004 NHANES (Centers for Disease Control and Prevention, 2004), which was 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;

Table 1. Percentages of missing data in the NHANES, 2003–2004

Partition†	*Variable*	*% missing, full data (n = 9041)*	*% missing, subset with HHINC observed (n = 5400)*
W	HHINC ($1000 per year)	40.27	0
Z	Age (years)	0	0
	Gender	0	0
X	Education (years)	17.24	16.74
	BMI (kg m^{-2})	9.84	9.48
Y	SBP (mm Hg)	25.02	24.5
	DBP (mm Hg)	25.02	24.5

†Partition based on covariate missingness and subsample MAR.

(b) two socio-economic status measures, household income HHINC and years of education EDU;
(c) three other covariates, age (in years), gender and BMI (in kilograms per metre squared).

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

Some of the variables have missing values—see Table 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 that is lost by CC analysis but assume that the missing values are MAR. It is reasonable to assume the values to be 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. If these assumptions are correct, IL methods yield biased regression estimates. This motivates the new method SIL, which allows assumptions of missingness at random for some variables (SBP, DBP, EDU and BMI) and assumptions of missingness not at random for others (HHINC), in a sense that is defined precisely in Section 4.

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

3. Complete-case and ignorable likelihood methods

In this section, we consider the data with the structure in Table 2. Let $\{(z_i, w_i, y_i), i = 1, \ldots, n\}$ 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 and Y have missing values. Interest concerns the parameters ϕ of the distribution of Y given (Z, W), say $p(y_i|z_i, w_i, \phi)$.

The rows of Table 2 divide the cases into two patterns. Pattern 1 $(i = 1, \ldots, m)$ consists of CCs, 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} denote the 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

Table 2. General missing data structure for Section 2†

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

†'✓' denotes observed, '×' denotes at least one entry missing and '?' denotes observed or missing.

denote a vector of 1s of the same length as the vector v, and $\bar{u}_v$ to denote a vector of 0s and 1s of the same length as v for which at least one entry is 0. Then, for the cases i in Table 2, $R_{(w_i,y_i)}=u_{(w,y)}$ for the CCs 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 $p(w_i,y_i|z_i,\theta)$—the fully observed covariates can be treated as fixed (Little and Rubin (2002), section 11.4.) The IL is obtained by integrating the missing variables out of this joint distribution, and treating θ as the argument of the resulting density, i.e.

$$L_{\text{ign}}(\theta)=\text{constant}\prod_{i=1}^{n}p(w_{\text{obs},i},y_{\text{obs},i}|z_i,\theta), \tag{1}$$

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 maximum likelihood (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 likelihood (1) is that the data are MAR, i.e.

$$p(R_{w_i},R_{y_i}|z_i,w_i,y_i,\psi)=p(R_{w_i},R_{y_i}|z_i,w_{\text{obs},i},y_{\text{obs},i},\psi), \tag{2}$$

where ψ are parameters for the missing data mechanism. If, in addition, the parameters θ and ψ are distinct, inferences based on likelihood (1) are fully efficient, but missingness at random 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 CCs, namely

$$L_{\text{cc}}(\phi)=\text{constant}\prod_{i=1}^{m}p(y_i|w_i,z_i,R_{(w_i,y_i)}=u_{(w,y)};\phi). \tag{3}$$

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

$$p(R_{(w_i,y_i)}=u_{(w,y)}|z_i,w_i,y_i,\psi)=p(R_{(w_i,y_i)}=u_{(w,y)}|z_i,w_i,\psi)\qquad\text{for all } y_i. \tag{4}$$

This condition allows missingness to be not at random, since it can depend on the values of W which are sometimes missing. CC analysis works in this case because equation (4) implies that

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

Table 3. Missing data pattern of example 1†

Pattern	*Observation i*	z_i	w_i	y_i	$R_{(w_i,y_i)}$
1	$1,\ldots,m$	✓	✓	✓	(1,1)
2	$m+1,\ldots,n$	✓	×	✓	(0,1)

†'✓' denotes observed and '×' denotes missing.

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

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

and the component $L_{\text{rest}}(\phi, \gamma)$ is discarded. ML estimates based on $L_{\text{cc}}(\phi)$ are consistent and asymptotically normal, but they are not necessarily fully efficient, since $L_{\text{rest}}(\phi, \gamma)$ 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.

3.1. Example 1: missing data in a single covariate

Table 3 displays a special case of Table 2 where w_i and y_i are single variables, and the incomplete cases have w_i missing (denoted ×) but not y_i. Condition (2) for values MAR becomes

$$p\{R_{(w_i,y_i)} = (1,1) | z_i, w_i, y_i, \psi\} = p\{R_{(w_i,y_i)} = (1,1) | z_i, y_i, \psi\} \qquad \text{for all } w_i, \tag{5}$$

and condition (4) becomes

$$p\{R_{(w_i,y_i)} = (1,1) | z_i, w_i, y_i, \psi\} = p\{R_{(w_i,y_i)} = (1,1) | z_i, w_i, \psi\} \qquad \text{for all } y_i. \tag{6}$$

The choice between IL or CC rests on whether condition (5) or (6) is a better assumption for the missing data mechanism, i.e. 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, presented 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 condition (6).

We note that CC analysis is viewed with disfavour 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) and Chen *et al.* (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 in the incomplete cases.

4. Subsample ignorable likelihood methods—theory

We consider the missing data pattern in Table 4, in which another set of incomplete covariates X is added. The observations are grouped into three patterns: pattern 1 consists of the CCs ($R_{w_i} = u_w$; $R_{(x_i,y_i)} = u_{(x,y)}$), pattern 2 incomplete cases with W fully observed ($R_{w_i} = u_w$;

Table 4. General missing data structure for Section 3†

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

†'✓' denotes observed, '×' denotes at least one entry missing and '?' denotes observed or missing.

$R_{(x_i,y_i)}=\bar{u}_{(x,y)}$) 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 $p(y_i|z_i,w_i,x_i,\phi)$. We propose SIL, 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 SIL 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, i.e.

$$p(R_{w_i}=u_w|z_i,w_i,x_i,y_i,\psi_w)=p(R_{w_i}=u_w|z_i,w_i,x_i,\psi_w) \qquad \text{for all } y_i. \tag{7}$$

(b) *Subsample missingness at random of X and Y*: X and Y are MAR within the subsample of cases for which W is fully observed, i.e.

$$p(R_{(x_i,y_i)}|z_i,w_i,x_i,y_i,R_{w_i}=u_w;\psi_{xy\cdot w})=p(R_{(x_i,y_i)}|z_i,w_i,x_{\text{obs},i},y_{\text{obs},i},R_{w_i}=u_w;\psi_{xy\cdot w}) \quad \text{for all } x_{\text{mis},i},y_{\text{mis},i}. \tag{8}$$

To establish the validity of SIL under conditions (7) and (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 $R_{w_i}=u_w$, i.e. restricted to cases i with W fully observed:

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

where $\zeta=(\theta,\psi)$. By a direct application of Rubin's (1976) theory, under the subsample missingness at random condition (8), this likelihood factorizes as

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

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

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

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

$$p(x_i,y_i|w_i,z_i,R_{w_i}=u_w;\theta)=p(y_i|x_i,w_i,z_i,R_{w_i}=u_w;\theta)\,p(x_i|w_i,z_i,R_{w_i}=u_w;\theta).$$

By assumption (7), $p(y_i|x_i, w_i, z_i, R_{w_i} = u_w; \theta) = p(y_i|x_i, w_i, z_i, \phi)$, 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 (7) and (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.

The missing data mechanism that is defined by conditions (7) and (8) is suitable in empirical studies where it is natural to assume covariate-dependent missingness for some covariates and subsample missingness at random for others. For example, in the motivating example concerning the regression of blood pressure on socio-economic variables in Section 2, HHINC may be covariate dependent and the education and BMI values may be subsample MAR. In environmental health research, values of variables that are missing because they lie below the limit of detection are MNAR. If missing values exist for other variables and can be assumed to be MAR, then SIL on the subsample with measurements within the detection limit yields valid regression inference.

Generally, SIL 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 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 (7) and (8).

Assumptions (7) and (8) differ from the assumptions under which IL and CC methods are valid. Specifically, IL inference assumes that the data are MAR, i.e.

$$p(R_{w_i}, R_{(x_i,y_i)}|z_i, w_i, x_i, y_i, \psi) = p(R_{w_i}, R_{(x_i,y_i)}|z_i, w_{\text{obs},i}, x_{\text{obs},i}, y_{\text{obs},i}, \psi) \quad \text{for all } w_{\text{mis},i},\ x_{\text{mis},i},\ y_{\text{mis},i}. \tag{9}$$

This differs from conditions (7) and (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, i.e.

$$p(R_{w_i} = u_w, R_{(x_i,y_i)} = u_{(x,y)}|z_i, w_i, x_i, y_i, \psi) = p(R_{w_i} = u_w, R_{(x_i,y_i)} = u_{(x,y)}|z_i, w_i, x_i, \psi) \quad \text{for all } y_i. \tag{10}$$

This differs from assumption (8) in that missingness of (x_i, y_i) in condition (8) can depend on the observed components of y_i. If this is not so, then CC yields valid inferences but is less efficient than SIL, since SIL uses the data in pattern 2, which are discarded by CC.

4.1. Example 2: normal regression model with two incompletely observed covariates

Table 5 displays a special case of Table 4, 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 (7) and (8) in this special case yields

$$p(R_{w_i} = 1|z_i, w_i, x_i, y_i, \psi_w) = p(R_{w_i} = 1|z_i, w_i, x_i, \psi_w) \quad \text{for all } y_i, \tag{11}$$

$$p(R_{x_i} = 1|z_i, w_i, x_i, y_i, R_{w_i} = 1, \psi_{xy\cdot w}) = p(R_{x_i} = 1|z_i, w_i, y_i, R_{w_i} = 1, \psi_{xy\cdot w}) \quad \text{for all } x_i. \tag{12}$$

Under this mechanism, SIL 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).

Table 5. Missing data structure for example 2†

Pattern	*Observation i*	z_i	w_i	x_i	y_i	R_{w_i}	R_{x_i}
1	$1,\ldots,m$	✓	✓	✓	✓	1	1
2	$m+1,\ldots,m+r$	✓	✓	×	✓	1	0
3	$m+r+1,\ldots,n$	✓	×	✓	✓	0	1

†'✓' denotes observed and '×' denotes missing.

5. Simulation study

As a numerical illustration of the theory in Section 4, we simulate data for the pattern of example 2, under a variety of missing data mechanisms. For each of 1000 replications, 5000 observations (z_i, w_i, x_i, y_i), $i=1,\ldots,5000$, 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

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

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

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

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 0; IL is valid if $\alpha_w^{(w)}$, $\alpha_x^{(w)}$ and $\alpha_x^{(x)}$ are 0; SIL is valid if $\alpha_y^{(w)}$ and $\alpha_x^{(x)}$ are 0. Four missing data mechanisms were created by using different sets of values for the regression coefficients such that, in mechanism I, all three methods (CC, IL and SIL) are consistent, whereas, in mechanisms II, III and IV, just one of the three methods is valid. The simulation set-up is summarized in Table 6.

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

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

(a) CC analysis, using ordinary least squares;
(b) IML for the whole data set;
(c) SIML, IML for the subsample with W observed;
(d) BD, least squares estimates from the regression before deletion, as a benchmark method.

For each method, Table 7 summarizes the root-mean-squared errors RMSE of estimates of all the regression coefficients, and Tables 8 and 9 report respectively the empirical bias and RMSE of estimates of the individual regression coefficients. Results in italics reflect situations

Table 6. Missing data mechanisms generated in the simulations†

Mechanism	$\alpha_0^{(w)}$	$\alpha_z^{(w)}$	$\alpha_w^{(w)}$	$\alpha_x^{(w)}$	$\alpha_y^{(w)}$	$\alpha_0^{(x)}$	$\alpha_z^{(x)}$	$\alpha_w^{(x)}$	$\alpha_x^{(x)}$	$\alpha_y^{(x)}$
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, SIML valid	−1	1	1	1	0	−2	1	1	0	1

†Missing values of W and X are generated on the basis of the following logistic models:

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

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

The four missing data mechanisms are as follows: I, missingness of $W=f(Z)$, missingness of $X=f(Z|W$ observed) and all four methods are valid; II, missingness of $W=f(Z,W,X)$, missingness of $X=f(Z,W,X|W$ observed) and only CC analysis is valid; III, missingness of $W=f(Z)$, missingness of $X=f(Z,W|$ observed) and only IML is valid; IV, missingness of $W=f(Z,W,Y)$, missingness of $X=f(Z,W,Y|W$ observed) and only SIML is valid.

Table 7. Summary RMSEs of estimated regression coefficients for BD, CCs, IML and SIML, under four missing data mechanisms†

Method	*RMSE× 1000 for the following values of ρ and mechanisms:*							
	ρ=0				*ρ=0.8*			
	I	*II*	*III*	*IV*	*I*	*II*	*III*	*IV*
BD	*27*	*28*	*28*	*27*	*50*	*46*	*50*	*46*
CC	*45*	*44*	553	322	*86*	*71*	426	246
IML	*37*	231	*36*	116	*58*	96	*53*	90
SIML	*42*	133	360	*49*	*70*	80	319	*69*

†The four missing data mechanisms are as follows: I, missingness of $W=f(Z)$, missingness of $X=f(Z|W$ observed) and all four methods are valid; II, missingness of $W=f(Z,W,X)$, missingness of $X=f(Z,W,X|W$ observed) and only CC analysis is valid; III, missingness of $W=f(Z)$, missingness of $X=f(Z,W|W$ observed) and only IML is valid; IV, missingness of $W=f(Z,W,Y)$, missingness of $X=f(Z,W,Y|W$ observed) and only SIML is valid. The RMSE-estimates are $1000\sqrt{E(\|\beta_r-\beta_{\text{TRUE}}\|^2)}$, with r denoting the rth repetition. Values in italics are for methods that are consistent for the mechanism generating the data.

where the method is consistent on the basis of the theory of Section 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 SIML lie between those for CC analysis and IML for mechanisms I, II and III, where one or both of CC analysis and IML are consistent—both CC and IML in mechanism I, CC analysis in mechanism II and IML in mechanism III. This finding reflects the fact that SIML is a hybrid of CC analysis and IML, sharing features of both methods. In mechanism IV, SIML is consistent but CC analysis and IML are inconsistent, and in this case SIML 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 little more detail.

For mechanism I, all three methods yield consistent estimates, IML is best since it makes full use of the data, CC analysis is the worst since it discards the most information and SIML lies between CC analysis and IML, since it retains some incomplete cases and drops others.

Table 8. Empirical bias for individual regression coefficients under four missing data mechanisms (1000 replications)

Method	*Bias× 1000 for the following mechanisms:*															
	Mechanism I				*Mechanism II*				*Mechanism III*				*Mechanism IV*			
	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
$\rho = 0$																
BD	*1*	*−1*	*1*	*1*	*0*	*2*	*1*	*−1*	*−4*	*2*	*−2*	*0*	*−3*	*3*	*1*	*2*
CC	*1*	*0*	*1*	*2*	*0*	*2*	*3*	*−1*	−458	−229	−171	−111	−262	−121	−123	−58
IML	*1*	*−1*	*2*	*1*	203	64	41	72	*−2*	*2*	*−3*	*2*	95	33	6	46
SIML	*2*	*1*	*2*	*2*	112	39	40	18	−292	−169	−88	−81	*−22*	*−10*	*−13*	*−4*
$\rho = 0.8$																
BD	*0*	*0*	*3*	*−3*	*0*	*−1*	*3*	*−1*	*2*	*−1*	*0*	*−2*	*0*	*−2*	*0*	*1*
CC	*−2*	*−5*	*5*	*0*	*0*	*−3*	*2*	*2*	−380	−137	−96	−71	−212	−75	−63	−32
IML	*−1*	*−5*	*6*	*−1*	72	20	5	27	*2*	*1*	*0*	*−3*	54	23	−11	39
SIML	*−1*	*−6*	*5*	*0*	42	9	17	2	−278	−119	−58	−61	*−16*	*−12*	*−6*	*2*

Table 9. RMSEs for individual regression coefficients under four missing data mechanisms (1000 replications)

Method	*RMSE × 1000 for the following mechanisms:*															
	Mechanism I				*Mechanism II*				*Mechanism III*				*Mechanism IV*			
	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
ρ = 0																
BD	*14*	*13*	*14*	*14*	*14*	*15*	*14*	*14*	*14*	*13*	*15*	*14*	*14*	*14*	*14*	*13*
CC	*21*	*22*	*26*	*22*	*25*	*20*	*21*	*20*	459	230	172	113	263	123	125	62
IML	*15*	*18*	*21*	*19*	204	67	45	74	*20*	*18*	*19*	*16*	97	37	18	49
SIML	*19*	*21*	*22*	*21*	114	43	45	26	293	170	90	82	*32*	*22*	*23*	*19*
ρ = 0.8																
BD	*15*	*27*	*27*	*28*	*15*	*26*	*24*	*25*	*14*	*25*	*27*	*30*	*14*	*26*	*22*	*27*
CC	*24*	*49*	*51*	*42*	*28*	*38*	*40*	*34*	381	141	101	80	214	83	73	50
IML	*17*	*30*	*36*	*30*	73	36	33	39	*16*	*28*	*27*	*32*	57	37	33	51
SIML	*20*	*39*	*39*	*39*	48	34	40	35	279	123	65	71	*28*	*37*	*35*	*38*

For mechanism II, CC analysis is valid and in general has the lowest RMSEs, whereas both IML and SSIML are biased, with SIML having RMSEs lying between those of CC analysis and IML. However, for $\rho = 0.8$, SIML and IML yield comparable or even smaller RMSEs than CC analysis 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 analysis and SIML lead to biased estimates, as shown in Table 7, with SIML being better than CC analysis since it incorporates features of IML as a method.

In mechanism IV, SIML is valid and CC analysis and IML are biased. The RMSEs from SIML are generally the smallest, except that IML yields a smaller RMSE than SIML 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 sizable sample size. This suggests a degree of robustness for IML, which has the virtue of retaining all the data.

6. Application to motivating example

We now apply the proposed method to the NHANES (2003–2004) data that were presented in Section 2. Two blood pressure measurements, systolic blood pressure SBP and diastolic blood pressure DBP, are regressed on household income HHINC, in dollars per year, and years of education EDU, adjusting for age (in years), gender and BMI, in kilograms per metre squared. 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 *versus* less than high school.

Age and gender are fully observed, whereas 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 we assume subsample missingness at random for other variables:

(a) missingness of BMI and blood pressure measurements is probably completely at random owing to missing visits;
(b) with income observed, it is reasonable to assume values MAR for education because income and education are correlated (Tolley and Olson, 1971).

With these two plausible assumptions, SIL on the subsample with household income observed yields consistent estimates of the regression, whereas 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, SIL is preferred over CC analysis since it uses more information in the incomplete cases than does CC analysis. For simplicity, we ignore the design features (weighting and clustering, etc.) of the NHANES. For the SIL method, we use IVEware to multiply-impute missing values in the subsample with household income observed, and then use SAS software (SAS Institute, 2010) to perform the regression analyses and to combine results from individual imputed data sets. We denote this method SIMI. For the IL method, we use IVEware to multiply-impute the full sample, and we use SAS software for regression analyses and combining the results. We denote this method IMI. The results of CC analysis, SIMI analysis and IMI are shown in Table 10. 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 the method of analysis. For education, SIMI and CC analysis yield similar and stronger effects

Table 10. Estimates of the effect of socio-economic status on blood pressure (NHANES, 2003–2004)

	CC analysis			*IMI analysis*			*SIMI analysis*		
	Estimate	*Standard error*	*p-value*	*Estimate*	*Standard error*	*p-value*	*Estimate*	*Standard error*	*p-value*
SBP									
Intercept	87.80	1.16	<0.0001	89.28	1.06	<0.0001	87.53	1.35	<0.0001
HHINC ($100000)	−0.84	0.97	0.3907	−0.84	1.11	0.4574	−0.88	0.94	0.3482
EDU (years)	−2.30	0.57	<0.0001	−2.06	0.44	<0.0001	−2.38	0.55	<0.0001
AGE (years)	0.49	0.01	<0.0001	0.50	0.01	<0.0001	0.50	0.01	<0.0001
Female	3.31	0.48	<0.0001	2.78	0.44	<0.0001	3.15	0.46	<0.0001
BMI (kg m^{-2})	0.46	0.04	<0.0001	0.41	0.03	<0.0001	0.47	0.04	<0.0001
DBP									
Intercept	45.46	1.06	<0.0001	46.94	1.00	<0.0001	45.46	1.19	<0.0001
HHINC ($100000)	2.97	0.89	0.0008	2.82	0.87	0.0026	2.83	0.97	0.0050
EDU (years)	4.86	0.52	<0.0001	4.06	0.43	<0.0001	4.95	0.52	<0.0001
AGE (years)	0.12	0.01	<0.0001	0.11	0.01	<0.0001	0.11	0.01	<0.0001
Female	1.81	0.44	<0.0001	1.83	0.36	<0.0001	1.86	0.42	<0.0001
BMI (kg m^{-2})	0.43	0.04	<0.0001	0.40	0.03	<0.0001	0.44	0.04	<0.0001

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 SIMI over CC analysis should result in a gain in efficiency for SIMI in this situation, although CC analysis and SIMI have similar estimated standard errors for this particular sample.

7. Discussion

The idea behind SIL, to apply an analysis that assumes values MAR to a subsample of the data that is complete on a subset of the covariates, is both simple and powerful. SIL analysis has the following strengths:

(a) it is easy to implement, since existing software for doing MAR value analyses is all that is required, and this software is now widely available for many common models;
(b) it avoids discarding all incomplete cases, thus alleviating one of the drawbacks of CC analysis;
(c) it applies to a broad class of univariate and multivariate regression models, including multivariate linear regression, generalized linear models and generalized linear mixed models;
(d) the method works for a class of missing data mechanisms, defined by conditions (7) and (8), where both IL and CC methods fail to give consistent estimates.

This extends the class of models for data MNAR that can be handled by a selective use of MAR data methods and allows combinations of MAR and MNAR data mechanisms for different variables in the data set.

In another analysis which drops a subset of incomplete cases, Von Hippel (2007) applied a multiple-imputation analysis with data MAR in the regression setting, where a univariate

outcome Y has missing values, and then applied the final regression analysis to the subsample of cases with Y observed, i.e. 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 data model and hence is much less general than SIL and does not generate a method that is consistent for a missingness not at random mechanism.

The general theoretical rationale of SIL is partial likelihood (Cox, 1972). This involves a potential loss of efficiency relative to full modelling, 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, software for full MNAR data models is not widely available.

An important topic is how much efficiency is lost by SIL relative to full likelihood methods. SIL 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 analysis 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 SIL method is in fact full ML for a particular normal pattern–set mixture model (Little, 1993). This aspect of SIL methods will be the subject of future work.

The form of IL method in SIL is left unspecified in this paper where possible, for increased generality. As noted, options for IL include 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 SIL is presented here in the context of likelihood-based analyses, but it also applies to non-likelihood analyses that are valid under the assumption of data MAR. 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, which is also valid under data MAR, without affecting the validity of the method under the stated assumptions (7) and (8).

From a practitioner's viewpoint, the main challenge in applying SIL is deciding which covariates belong in the set W and which belong in the set X, i.e. which covariates are used to create the subsample for the missingness-at-random analysis. The choice is guided by the basic assumptions (7) and (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, e.g. 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 analysis, IL and weighted generalized estimating equations. When faced with missing data, assumptions are inevitable, and they need to be as reasonable and as 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 analysis, IL and SIL for subsamples judiciously chosen on the basis of assumptions (7) and (8), to assess sensitivity of key inferences to alternative assumptions about the missing data mechanism.

Acknowledgements

We thank the Joint Editor, the Associate Editor and two referees for their thoughtful and constructive comments which greatly improved the paper. The authors are listed in alphabetical order.

References

Centers for Disease Control and Prevention (2004) *National Health and Nutrition Examination Survey Data*. Hyattsville: 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. *J. Multiv. Anal.*, **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. Ass.*, **102**, 1309–1317.

Colhoun, H., Hemingway, H. and Poulter, N. R. (1998) Socio-economic status and blood pressure: an overview analysis. *J. Hum. Hypertens.*, **12**, 91–110.

Cox, D. R. (1972) Partial likelihood. *Biometrika*, **62**, 269–276.

Gilks, W. R., Thomas, A. and Spiegelhalter, D. J. (1994) A language and program for complex Bayesian modelling. *Statistician*, **43**, 169–177.

Glynn, R. J. and Laird, N. M. (1986) Regression estimates and missing data: complete-case analysis. *Technical Report*. Department of Biostatistics, Harvard School of Public Health, Boston.

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

Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. (2002) Bayesian methods for generalized linear models with covariates missing at random. *Can. J. Statist.*, **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. Ass.*, **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 non-ignorable. *J. R. Statist. Soc.* B, **61**, 173–190.

Little, R. J. A. (1979) Maximum likelihood inference for multiple regression with missing values: a simulation study. *J. R. Statist. Soc.* B, **41**, 76–87.

Little, R. J. A. (1992) Regression with missing X's: a review. *J. Am. Statist. Ass.*, **87**, 1127–1137.

Little, R. J. A. (1993) Pattern-mixture model for multivariate incomplete data. *J. Am. Statist. Ass.*, **88**, 125–134.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. Hoboken: Wiley.

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

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. *Publ. Hlth Rep.*, **112**, 753–763.

Mackenbach, J. P. (1994) The epidemiologic transition theory. *J. Epidem. Commty Hlth*, **48**, 329–331.

Raghunathan, T., Lepkowski, J., VanHoewyk, M. and Solenberger, P. (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. *Surv. Methodol.*, **27**, 85–95.

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

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

SAS Institute (2010) *Statistical Analysis with SAS/STAT® Software*. Cary: SAS Institute.

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. Ass.*, **101**, 924–933.

Tolley, G. S. and Olson, E. (1971) The interdependence between income and education. *J. Polit. Econ.*, **79**, 460–480.

Von Hippel, P. T. (2007) Regression with missing Ys: an improved strategy for analyzing multiply imputed data. *Sociol. Methodol.*, **37**, 83–117.