A Pseudo-Bayesian Shrinkage Approach to Regression with Missing Covariates

Nanhua Zhang[1],* and Roderick J. Little[2],**

[1]Department of Epidemiology & Biostatistics, College of Public Health, University of South Florida, Tampa, Florida 33612–3085, U.S.A.
[2]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan 48109–2029, U.S.A.
*email: nzhang1@health.usf.edu
**email: rlittle@umich.edu

SUMMARY.   We consider the linear 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 (i) complete-case analysis (CC), which discards the incomplete cases, and (ii) ignorable likelihood methods, which base inference on the likelihood based on the observed data, assuming the missing data are missing at random (Rubin, 1976b), and (iii) 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 (i) the regression coefficient of W is zero or (ii) 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 covariate is also discussed.

KEY WORDS: Complete-case analysis; Drop variables analysis; Gibbs sampling; Nonignorable modeling; Shrinkage; Variable selection.

1. Introduction

We consider multivariate regression with missing covariates, with the structure of the data displayed in Figure 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 Y on Z, adjusting for W. The incomplete cases have very little information for the coefficient of W (Little, 1992), and because 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, Lipsitz, and Chen (1999), Ibrahim, Chen, and Lipsitz (2002), Ibrahim et al. (2005), and Chen et al. (2008). Three common approaches are:

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

(ii) Ignorable likelihood (IL) methods, 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;

(iii) Nonignorable modeling, 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 (MAR), 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 a nonignorable modeling method, but such methods are vulnerable to misspecification of the missing data mechanism, and suffer from problems with identifying the parameters (see, e.g.,

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

Key: √ denotes observed, x denotes missing

Figure 1. Missing data structure in Section 1.

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

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

Figure 2. Missing data structure for Section 3.

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, because it may be worth accepting a small amount of bias in the regression estimates 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 nonsignificant 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 pseudo-Bayesian (PB) 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 2 presents a motivating example using data from two Eastern Cooperative Oncology Group clinical trials. Section 3 reviews properties of CC and DV, in a slightly more general regression setting. In Section 4, we propose a PB 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 5 presents some simulations that demonstrate attractive properties of the proposed method, and in section 6 we apply the proposed method to a liver cancer data set. Extensions to more than one missing regressors are discussed in Section 7.

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, Cnaan, and Simson, 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, Chen, and Lipsitz, 1999; Huang, Chen, and Ibrahim, 2005; Chen, Zeng, and Ibrahim, 2007; Das, Maiti, and Pradhan, 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; and
(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 MAR (Rubin, 1976b; Little and Rubin, 2002). We propose a PB approach for this problem, which compromises between the CC and DV estimates.

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

3. Complete Case and Drop Variable Analyses

In this section, we consider the data with the structure in Figure 2. Let $\{(z_i, w_i, y_i), i = 1, \dots n\}$ 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 $p(y_i | z_i, w_i, \phi)$.

The rows of Figure 2 divide the cases into two patterns. Pattern 1 $(i = 1, \dots, m)$ 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, $R_{w_i} = u_w \equiv (1, \ldots, 1)$, 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:

$$L_{\text{cc}}(\phi) = \text{const.} \times \prod_{i=1}^{m} p\left(y_i | w_i, z_i, R_{w_i} = u_w; \phi\right), \qquad (1)$$

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, that is:

$$
\begin{aligned}
p(R_{w_i} &= u_w | z_i, w_i, y_i, \psi)) \\
&= p(R_{w_i} = u_w | z_i, w_i, \psi)) \quad \text{for all } y_i, \qquad (2)
\end{aligned}
$$

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

$$p(y_i | w_i, z_i, R_{w_i} = u_w, \phi) = p(y_i | w_i, z_i, \phi),$$

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, γ) can be written as

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

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

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:

$$L_{\text{DV}}(\phi_z) = \text{const.} \times \prod_{i=1}^{n} p\left[y_i | z_i, \beta_w = (0, \ldots, 0); \phi_z\right]. \quad (3)$$

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 $\beta_w \neq (0, \ldots, 0)$ when W and Z are not associated. This fact will be exploited in the proposed method, which we now describe.

4. Pseudo-Bayesian Shrinkage Method for Regression with Missing Covariates

4.1 *Motivation*

In this section, we consider the data structure in Figure 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:

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

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:

(1) $\beta_w = 0$;
(2) $\rho_{wz^*} \equiv \text{cov}(W, Z^*) = 0$, 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, because 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.

4.2 *Modeling*

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

$$\beta_w | J \sim J\delta(0) + (1 - J)N\left(0, \tau_w^2\right), \qquad (4)$$

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

$$\Pr(J = 0) = 1 - \Pr(J = 1) = \pi_0. \qquad (5)$$

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 (4) in the full prior distribution, we use a multivariate normal prior

$$\beta | J \sim N(0, D_J D_J), \qquad (6)$$

with

$$D_J \equiv diag[1, (1 - J)\tau_w, \tau_z]. \qquad (7)$$

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

$$\sigma^2 | J \sim IG(\nu_J / 2, \nu_J \lambda_J / 2). \qquad (8)$$

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.

4.3 A Pragmatic Choice of π_0

As we can see from Section 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 CC. 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:

$$\pi_0 = f(|\rho_{wz^*}|) = |\rho_{wz^*}|. \tag{9}$$

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

$$L(\mu_{wz^*}, \Sigma_{wz^*}) = \prod_{i=1}^{m} f(w_i, z_i^*|\mu_{wz^*}, \Sigma_{wz^*}). \tag{10}$$

4.4 Estimation

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

$$\beta = \left(1, \beta_w, \beta_z^T\right)^T, \quad X_{CC} = \begin{bmatrix} 1 & w_1 & z_1^T \\ 1 & w_2 & z_2^T \\ & \cdots & \\ 1 & w_m & z_m^T \end{bmatrix},$$

$$X_{DV} = \begin{bmatrix} 1 & z_1^T \\ 1 & z_2^T \\ \cdots & \\ 1 & z_n^T \end{bmatrix}, \quad Y_{CC} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_m \end{bmatrix}, \quad Y_{DV} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix}.$$

Also, let $\hat{\beta}_{DV,LS}$ be the least-square estimate based on DV analysis, and $\hat{\beta}_{CC,LS}$ be the least-square estimate based on CC. Accordingly, the estimated residual variances for DV and CC analysis are denoted as $\hat{\sigma}_{DV,LS}^2$ and $\hat{\sigma}_{CC,LS}^2$.

The chain is initialized at a starting value $\beta^0, \sigma^{2(0)}$, and J^0. A reasonable starting value for J^0 is 0, which is a complete-case scenario, and therefore the corresponding starting values for $\beta^0, \sigma^{2(0)}$ are $\hat{\beta}_{CC,LS}$ and $\hat{\sigma}_{CC,LS}$. First, β^k and $\sigma^{2(k)}$ can be sampled in the following way:

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

$$\beta^k = \left[\beta_0^k, \beta_w^k, \left(\beta_z^k\right)^T\right]^T$$
$$\sim N\left[A_{CC,J^{k-1}}(\sigma^{k-1})^{-2} X_{CC}^T X_{CC} \hat{\beta}_{CC,LS}, A_{CC,J^{k-1}}\right], \tag{11}$$

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

from

$$\sigma^{2(k)} \sim f(\sigma^{2(k)}|\beta^k, J^{k-1})$$
$$= IG\left(\frac{n + \nu_{J^{k-1}}}{2}, \frac{|Y_{CC} - X_{CC}\beta^k|^2 + \nu_{J^{k-1}}\lambda_{J^{k-1}}}{2}\right). \tag{12}$$

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

$$\beta^k = [\beta_0^k, 0, (\beta_z^k)^T]^T \text{ with}$$

$$\left[\beta_0^k, \left(\beta_z^k\right)^T\right]^T$$
$$\sim N\left[A_{DV,J^{k-1}}(\sigma^{2(k-1)})^{-1} X_{DV}^T X_{DV} \hat{\beta}_{DV,LS}, A_{DV,J^{k-1}}\right], \tag{13}$$

where $A_{DV,J^{k-1}} = [(\sigma^{2(k-1)})^{-1} X_{DV}^T X_{DV} + D_{J^{k-1}}^{-1} D_{J^{k-1}}^{-1}]^{-1}$ and $D_J^{-1} = diag[1, \tau_z^{-1}]$; and $\sigma^{2(k)}$ is obtained by sampling from

$$\sigma^{2(k)} \sim f(\sigma^{2(k)}|\beta^k, J^{k-1})$$
$$= IG\left[\frac{n + \nu_{J^{k-1}}}{2}, \frac{\left|Y_{DV} - X_{DV}\left(\beta_0^k, \left(\beta_z^k\right)^T\right)^T\right|^2 + \nu_{J^{k-1}}\lambda_{J^{k-1}}}{2}\right]. \tag{14}$$

Next, $\rho_{wz^*}^k$ 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

$$\Pr\left(J^k = 1|\beta^k, \sigma^{2(k)}, \rho_{wz^*}^k\right) = \frac{r}{r+s}, \tag{15}$$

with $r = f(\beta^k|J^k = 1)(1 - |\rho_{wz^*}^k|)$ and $s = f(\beta^k|J^k = 0)|\rho_{wz^*}^k|$.

Note that, when $J^{k-1} = 0$, the conditional distribution of $\beta^k, \sigma^{2(k)}$ are based on the complete-case likelihood, which is a partial likelihood. Because 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.

4.5 Posterior Probability that $J = 0$, π_1

The posterior probability of $J = 0$, π_1, namely, the posterior probability of including the incomplete variable W in the regression model and using CC, 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.

5. Simulation Studies

In this section we describe simulations that illustrate the properties of the PB approach in Section 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 $i = 1, 2, \ldots, 100$. 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 four missing data mechanisms:

(1) MCAR: $\Pr(M_{w_i} = 1 | w_i, z_{1i}, z_{2i}, y_i) = 0.25$;
(2) MNAR: $\Pr(M_{w_i} = 1 | w_i, z_{1i}, z_{2i}, y_i) = \text{expit}(w_i - 1)$;
(3) MAR1: $\Pr(M_{w_i} = 1 | w_i, z_{1i}, z_{2i}, y_i) =$
   $\text{expit}(z_{1i} + z_{2i} - 1)$;
(4) 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 ($\rho = 0, 0.3, 0.8$) and two regression coefficients for W ($a = 0, 1$), yielding 24 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 assuming MAR;
(3) CC: Complete-case analysis;
(4) DV: dropping the missing covariate W; and
(5) PB: the proposed pseudo-Bayesian shrinkage method between CC and DV.

We report the ratios of root mean square errors (RMSEs) of IL, CC, DV, and PB method 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 1, 2, and 3. Results are based on 1000 repetitions for each simulation condition. Table 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 three missing data mechanisms because missingness does not depend on the outcome, but biased for the fourth missing data mechanism because missingness of W is dependent on the outcome. There is some loss of information because the incomplete cases are dropped from the analysis. IL is consistent and efficient for missing data mechanism I, III, and IV because all are MAR. DV is valid when $a = 0$ or $\rho = 0$, and in these cases the PB estimates are close to DV; when $a \neq 0$ and $\rho \neq 0$, PB estimates yield 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 MAR mechanisms I, III and IV but has larger RMSE than PB estimates when the data are MNAR.

6. Application to a Liver Cancer Study

We now apply the proposed method to the liver cancer data presented in Section 2. We regress the baseline number of CNTs on four baseline characteristics: BMI, age in year, associated jaundice (yes, no), and TSD of the disease (in weeks).

To be consistent with Chen et al. (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 $TSD^{1/2}$ 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 et al. (2007), TSD is assumed to be MAR. However, it is likely that TSD is not MAR because 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 correlations between TSD and BMI, age, jaundice are –0.020, 0.013, and 0.009, respectively. The correlation between TSD and the linear combination of BMI, age, and jaundice weighted by the regression coefficients using complete cases is –0.002.

Table 4 shows the results of applying the PB shrinkage method, CC and DV. We ran 10,000 iterations and obtain draws of the posterior estimates of the regression coefficients. The posterior probability of including TSD is 0.0153, indicating that the PB method favors dropping TSD from the regression and using full sample. This is not surprising, because 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:

$$df = n \times (1 - \hat{\pi}_1) - k - \hat{\pi}_1, \qquad (16)$$

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 4, both CC and PB methods show that age and jaundice are related to the number of CNTs, while BMI and TSD are not significant. However, the PB method yields smaller standard error for the regression estimate; therefore, the effect of age and jaundice are stronger than CC. Because the posterior probability of keeping TSD in the modeling is very small ($\hat{\pi}_1 = 0.0153$), the PB method is very similar to the regression without TSD.

7. Discussion

We have described a PB shrinkage method for regression analysis with a missing covariate, which is a compromise between CC 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. 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 MNAR, 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 IL methods are preferable in that case, because they are asymptotically efficient; the IL also works

Table 1
RMSE ratios for individual regression coefficients under four 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
		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
		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
		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
		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
		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
		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

*MD 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) MAR1, $\Pr(M_{w_i}=1|w_i,z_{1i},z_{2i},y_i)=\text{expit}(z_{1i}+z_{2i}-1)$; and (iv) MAR2, $\Pr(M_{w_i}=1|w_i,z_{1i},z_{2i},y_i)=\text{expit}(y_i-2)$.

†$\hat{\pi}_1$: the posterior probability of including the incomplete covariate.

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

Table 2

95% confidence coverage for individual regression coefficients under four missing data mechanisms (1000 replications)

ρ	a	MD* mechanism	β₀					β_w					β_z1					β₂₂				
			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
		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
		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	94.3	94.1	94.6	95.4	93.9	94.6	94.8	94.6	95.4
		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	95.4	92.9	93.7	87.8	94.8
		MAR	94.3	93.4	94.2	94.7	94.3	95.4	94.0	94.4	0.0	83.2	94.8	95.1	95.4	85.8	95.6	94.8	93.5	93.8	86.8	95.2
		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
		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
		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

*MD mechanisms: (i) MCAR, $\Pr(M_{w_i}=1|w_i,z_{1i},z_{2i},y_i)=0.25$; (ii) MNAR, $\Pr(M_{w_i}=1|w_i,z_{1i},z_{2i},y_i)=\mathrm{expit}(w_i-1)$; (iii) MAR1, $\Pr(M_{w_i}=1|w_i,z_{1i},z_{2i},y_i)=\mathrm{expit}(z_{1i}+z_{2i}-1)$; (iv) MAR2, $\Pr(M_{w_i}=1|w_i,z_{1i},z_{2i},y_i)=\mathrm{expit}(y_i-2)$.

**BD: before deletion; IL: ignorable maximum likelihood; CC: complete-case analysis; DV: dropping variable W; PB: pseudo-Bayesian analysis.

Table 3
Bias (Z-score) for individual regression coefficients under four 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
		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
		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
		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
		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
		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
		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

*MD 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) MAR1, $\Pr(M_{w_i} = 1|w_i, z_{1i}, z_{2i}, y_i) = \text{expit}(z_{1i} + z_{2i} - 1)$; (iv) MAR2, $\Pr(M_{w_i} = 1|w_i, z_{1i}, z_{2i}, y_i) = \text{expit}(y_i - 2)$.
**IL: ignorable maximum likelihood; CC: complete-case analysis; DV: dropping variable W; PB: pseudo-Bayesian analysis.

Table 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	<0.0001	2.601	0.401	<0.0001	2.545	0.369	<0.0001	2.503	0.370	<0.0001
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
$TSD^{1/2}$	−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.

well under slight violations of the MAR assumption (Little and Zhang, 2011). The proposed PB shrinkage method does yield biased estimates of the intercept and the regression coefficient of the incomplete covariate, W, as with any Bayesian method that puts positive prior probability for this coefficient on zero. Whether this bias is a serious limitation is debatable, but here our focus is on the potential recovery of information for the coefficient of Z, adjusting for W.

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 4.2. In this case, $\rho_{w_1 z}, \ldots, \rho_{w_d z}$ represents the d correlation coefficients between W and Z^*, and draws of the indicators $J_1, \ldots, J_d$ for whether the corresponding coefficients are zero are sampled in the estimation step. We suggest sampling $J_1, \ldots, J_d$ in a random order to get fast convergence of the chain. For the jth component of W, $Y_{\mathrm{CC},j}$ is defined to be the vector of outcomes corresponding to CC, 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; van Buuren and Oudshoorn, 2000) to impute missing values of Z.

There is a potential loss of efficiency of the PB 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.

References

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

Chen, Q., Zeng, D., and Ibrahim, J. G. (2007). Sieve maximum likelihood estimation for regression models with covariates missing at random. *Journal of the American Statistical Association* **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.

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.

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.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88,** 881–889.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement* **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., 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, Series B* **61,** 173–190.

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. *Journal of the American Statistical Association* **100,** 332–346.

IVEware. (2011). *Imputation and Variance Estimation Software*. Ann Arbor, M.I.: Institute of Social Research, University of Michigan. Available at http://www.isr.umich.edu/src/smp/ive/, accessed 20 October, 2011.

Little, R. J. A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association* **87,** 1127–1137.

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

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

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd edition. Hoboken, NJ: 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. (2011). Subsample ignorable likelihood for regression with missing data. *Journal of the Royal Statistical Society, Series C* **60,** 591–605.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83,** 1023–1032.

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

Rubin, D. B. (1976a). Comparing regressions when some predictor values are missing. *Technometrics* **18,** 201–205

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

van Buuren, S. and Oudshoorn, C. G. M. (2000). *Multivariate imputation by chained equations: MICE V1.0 user's manual*. TNO Report PG/VGZ/00.038. TNO Preventie en Gezondheid, Leiden, The Netherlands.