BIOMETRICS , 1–20
2011

Variable Selection for Multiply-Imputed Data with Application to Dioxin Exposure Study

University of Wisconsin-Madison

Department of Biostatistics and Medical Informatics

Technical Report #217

Qixuan Chen[1]* and Sijian Wang[2,3]**

[1]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA

[2]Department of Biostatistics & Medical Informatics, University of Wisconsin, Madison, Madison, WI 53792, USA

[3]Department of Statistics, University of Wisconsin, Madison, Madison, WI 53792, USA

*qc2138@columbia.edu; **swang@biostat.wisc.edu

SUMMARY: Multiple imputation (MI) is a commonly used technique for handling missing data in large-scale medical and public health studies. An important and longstanding statistical problem is how to conduct variable selection on the multiply-imputed data. If a variable selection method is applied to each imputed dataset separately, it may select different variables in different imputed datasets and make it difficult to interpret the final model or draw the scientific conclusions. In this paper, we propose two novel variable selection methods for the multiply-imputed data. Both methods jointly fit models on multiple imputed datasets and yield a consistent selection of variables across all imputed datasets. The first method is called MI-stepwise method, which is an extension of the stepwise selection method. It is implemented by first obtaining combined *p-values* using Rubin's rules for MI inference (Rubin 1987; Little and Rubin 2002) and then selecting variables base on the combined *p-values* in each step of selection. The second method is called MI-lasso method, which is an extension of the lasso method (Tibshrani, 1996). It treats the estimated regression coefficients of the same variable across all imputed datasets as a group, and applies the group lasso penalty (Yuan and Lin 2006) to yield a consistent variable selection across multiple imputed datasets. The proposed methods are demonstrated using simulation studies. We also apply the two methods to the University of Michigan Dioxin Exposure Study (UMDES) to identify important environmental exposure factors that are associated with serum dioxin concentrations.

KEY WORDS: exposure assessment; group lasso penalty; regularization; Rubin's rules; stepwise selection; variable selection.

1. Introduction

Dioxins are a class of chemical contaminants that are highly toxic and persist in the environment. Studies of highly exposed populations show that dioxins can cause excess cancer incidence, diabetes, immune system suppression, skin problems and much more (Passi et. al., 1981; Dalton et. al., 2001; Arisawa et. al., 2005; Baccarelli et. al., 2002; Zambon et. al., 2007). Since dioxins are very difficult to resolve once they enter human's body, it is important to investigate the site-specific exposure factors and pathways by which people accumulate body burdens of dioxins. This information can aid the development of regulations to reduce the risk of exposure.

The work of this article is motivated by analyzing The University of Michigan Dioxin Exposure Study (UMDES) data. The UMDES is the first population-based dioxin exposure study of residents living in the Midland, Michigan area, one of the largest and best characterized dioxin contamination sites in North America (Garabrant et al., 2009a, 2009b; Hedgeman et al., 2009). In the UMDES, adults older than 18 years of age who had lived in their current residence for at least 5 years were eligible to participate. Eligible subjects were selected from the populations of five counties in Michigan and invited to complete an interview, donate an 80-mL whole blood sample, have their household dust and soil sample collected. The study generated an unprecedented dataset with an extensive collection of information on demographics, health, residence, activities, work history, lifetime food consumption, current diet, and measurements of dioxins in participants' household dust, soil, and serum.

As often encountered in large epidemiological studies, the statistical analysis of the UMDES was hindered by missing data, which were caused by item nonresponse in survey questionnaire and missing dust and soil values from subjects who refused to provide samples. Although the nonresponse rate is small in each individual variable, the missing values disperse throughout the data in a haphazard pattern. Thus, ignoring missing data by deleting incomplete cases is wasteful of the costly collected data and can lead to a biased statistical inference. Alternatively, the multiple

imputation (MI) framework suggested by Rubin (1987) can be used to address the problem of missing data. MI refers to a procedure of replacing each missing value with plausible values multiple times to generate multiple complete datasets. MI has a practical advantage of allowing standard complete-data methods of analysis to be used, and the complete-data inferences from each imputed dataset can be combined to form one inference that properly reflects the uncertainty of imputation due to the missing data. Compared to maximum-likelihood estimates calculated directly from the incomplete data, MI is more attractive for the analysis of the incomplete data with the general missing pattern and for multiple-purpose analysis with many estimands like the UMDES data. Unless the rate of missing information is very high, in most situations only three to ten imputations are enough to yield an excellent efficiency (Rubin 1987). For the UMDES data, Olson et al. (2006) used a sequential regression imputation procedure (Raghunathan et al., 2001) with five imputations to generate values for missing items. Our analysis is based on these five imputed datasets, and our goal is to identify important environmental exposure factors that are associated with human population's serum dioxin concentrations in Midland. In statistics, this is a variable selection problem.

Variable selection has been extensively studied in statistical literature. Some classical methods including stepwise selection or best subset selection search for the best model based on either significance test or a certain information based criterion, such as Akaike information criterion (AIC, Akaike 1974) and Bayesian information criterion (BIC, Schwarz 1978). Penalized likelihood-based methods have drawn a lot of attention in recent literature, including lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), SICA (Lv and Fan, 2009), MCP (Zhang, 2010), group lasso (Yuan and Lin, 2006), CAP (Zhao, et. al. 2009), and grouping pursuit (Shen and Huang, 2010), among others. Bayesian variable selection strategies (George and McCulloch, 1993; George and Foster, 2000; Park and Casella, 2008) have also become popular in many applications. However, it may yield an inadequate performance if these variable selection methods are directly applied to the multiply-

imputed data. The main problem is on how to achieve the selection consistency across the multiple datasets generated by imputation. Specifically, if a variable selection method is applied to each individual imputed dataset separately, it will probably identify different important variables in each imputed dataset. This makes it difficult to produce the overall parameter estimates across all imputed datasets, and hence makes it difficult to interpret the model or draw scientific conclusions.

When MI is used to handle missing data, some naïve variable selection strategies are often used, such as conducting variable selection with complete-cases only, i.e., discarding cases with any missing values, or with a single imputed dataset. These naïve strategies are easy to implement but they ignore either the possible difference between the complete cases and incomplete cases or the uncertainty of imputation caused by missing information. To incorporate the missing-data uncertainty, one popular approach (Heymans et al., 2007) in epidemiological literature is to conduct variable selection in each imputed dataset separately, and a variable is claimed to be important if it is selected with a frequency greater than some subjective threshold (say, 60%). Obviously, this approach may identify different important variables with different thresholds, and there is no clear guide on how to choose a proper threshold. Two works in statistical literature have addressed the variable selection problem for multiply-imputed data. Yang, Belin, and Boscardin (2005) proposed a Bayesian variable selection method for multiply-imputed data, by applying Rubin's rules to synthesize different sets of Markov chain Monte Carlo estimates into a final summary. Wood et al. (2008) proposed a method which stacks the multiple imputed datasets and uses a weighting scheme to account for the percentage of missing values in each variable. Backward stepwise selection is then applied to the stacked dataset to identify the best model.

In this paper, we describe two novel variable selection methods for linear regression model with multiply-imputed data. The first method extends the stepwise selection method to multiply-imputed data by repeatedly applying Rubin's rules to select or remove variables based on combined *p-values* and produce combined inferences in each step. We call this method "MI-stepwise".

The second method extends the lasso method to multiply-imputed data by treating the estimated regression coefficients associated with the same variable across different imputed datasets as a group and selecting or removing the whole group together. We call this method "MI-lasso". Both methods incorporate the missing-data uncertainty in the variable selection procedure and yield a consistent selection across multiple imputed datasets. After models are selected, the overall estimation and inference can be produced by applying Rubin's rules.

The rest of the paper is organized as follows. In Section 2, we briefly describe the Rubin's rules for MI inference. We introduce our new methods: MI-stepwise and MI-lasso in Section 3 and 4, respectively. In Section 5, we demonstrate our methods using simulation studies. In Section 6, we apply our two methods to the UMDES data to identify important exposure factors that are associated with population's serum dioxin concentrations in Midland. We conclude the paper with Section 7.

2. Rubin's rules for MI inference

2.1 *Notation*

Suppose we have n subjects and p variables for each subject. Let Y_i be the outcome variable and X_{ij} be the j^{th} variable ($j = 1, \ldots, p$) for the i^{th} subject. Without loss of generality, we assume that Y_i's are centered to have zero mean and X_{ij}'s are standardized to have zero mean and unit standard deviation. Consider the following linear regression model:

$$Y_i = \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where β_j's are regression coefficients, and $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ are error terms. Since X_{ij}'s and Y_i's are centered, the intercept can be excluded from the model. We are interested in determining a smaller subset of variables that exhibits the strongest effects.

We assume that Y_i's are fully observed, but X_{ij} is partially observed for some j. The observation indicator for X_{ij} is denoted by R_{ij}; if $R_{ij} = 1$, X_{ij} is observed and, if $R_{ij} = 0$, X_{ij} is missing. The

missing values in the variables are imputed D times to generate D imputed datasets. Details of multiple imputation procedure can be found elsewhere (Meng, 1995; Raghunathan et al., 2001; Rubin, 1987, 1996; Schafer, 1997;). We denote each of the D imputed datasets as $(y_i; x_{d,i1}, \ldots, x_{d,ip})_{i=1}^n$, $d = 1, \ldots, D$, where $x_{d,ij}$ is the value of variable X_j for the i^{th} subject in the d^{th} imputed dataset. If $R_{ij} = 1$, we have $x_{1,ij} = \cdots = x_{D,ij} = x_{ij}$, and if $R_{ij} = 0$, $x_{d,ij}$ can take different values in each imputation.

2.2 *Rubin's rules*

With a given model structure, i.e., the variables included in the model are determined, the multiply-imputed data can be analyzed using Rubin's rules (Rubin, 1987; Little and Rubin, 2002). The general idea is briefly stated as follows. First, the parameters of interest are estimated based on each imputed dataset separately. Then the multiple estimates of the same parameter from all imputed datasets are combined into an overall estimate and the corresponding inference incorporates both within-imputation and between-imputation variation. To be specific in the linear regression model, let $\hat{\beta}_{d,j}$ and $V_{d,j}$, $d = 1, \ldots, D$, be D regression coefficient estimates and their associated estimated variances, which are calculated from D imputed datasets under Model (1). The final combined estimate of β_j is the average $\bar{\beta}_j = \frac{1}{D}\sum_{d=1}^{D} \hat{\beta}_{d,j}$. The variance of $\bar{\beta}_j$ has two components: the average within-imputation variance which can be estimated by $\bar{V}_j = \frac{1}{D}\sum_{d=1}^{D} V_{d,j}$ and the between-imputation variance which can be estimated by $B_j = \frac{1}{D-1}\sum_{d=1}^{D}(\hat{\beta}_{d,j} - \bar{\beta}_j)^2$. The combined variance associated with $\bar{\beta}_j$ is then defined as $T_j = \bar{V}_j + \frac{D+1}{D}B_j$. With a large sample size,

$$(\bar{\beta}_j - \beta_j)T_j^{-1/2} \sim t_\nu, \tag{2}$$

where the degrees of freedom $\nu = (D-1)(1+\frac{1}{D+1}\frac{\bar{V}_j}{B_j})^2$ (Rubin, 1987; Rubin and Schenker, 1986). We use p_j to denote the combined *p-value* of the *t*-test associated with the j^{th} variable in (2). Notice that if variables included in the model for each imputed dataset are different, it is not clear how to apply the Rubin's rules. This again illustrates that it is desired to have a variable selection method which yields a consistent selection across all imputed datasets.

3. MI-stepwise method

Stepwise variable selection methods (Efroymson, 1960; Kutner, et. al., 2004) based on significance tests are widely used in medical and public health research. Starting with a certain model, the stepwise variable selection tests the significance of each variable that is not in the model one by one. The variable that is most significantly associated with the outcome is selected into the model. After a new variable is added, the significance of the variables already in the model is checked. Variables that become insignificant as the new variable is added to the model are sequentially removed from the model. This procedure terminates when no more variables not in the model can be included and no more variables in the model can be removed. When analyzing multiply-imputed data, if the stepwise selection method is applied to each imputed dataset separately, in each step of selection, it is likely that different variables are selected or removed in different imputed datasets, and the final model based on each imputed dataset may select different variables. The reason for this inconsistency of selection is that, in each step of selection, the significance of each variable is evaluated in each imputed dataset *separately*. If we infer the significance for each variable *jointly* across imputed datasets and assign a common *p-value* for each variable, the action for each variable (select, remove, or stay in the model) will be the same in all imputed datasets, and hence the selection consistency across all imputed datasets can be achieved. This intuition motivates our MI-stepwise method whose selection procedures are *jointly* conducted across all imputed datasets. In each selection step, the Rubin's rules are used to obtain the combined *p-value* for each variable first, and then the action on each variable is determined based on this combined *p-value*. Without loss of generality, we present the detailed procedure for our MI-forward-stepwise selection method as follows. The MI-backward-stepwise selection method can be conducted in a similar way.

Step 0: Determine α_1, the significance level for including a variable in the model, and α_2, the significance level for removing a variable from the model. Specify the initial model M_0. Set $t = 0$.

Step 1: Let $t = t + 1$. For each variable X that is not included in M_{t-1}, fit D regressions with the model $\{M_{t-1}, X\}$ on D imputed datasets. Calculate the combined *p-value* for the newly added variable X in each model. Denote X_a be the variable with the smallest combined *p-value* p_a. If $p_a \leqslant \alpha_1$, update the model M_t to be $\{M_{t-1}, X_a\}$; otherwise, the procedure terminates.

Step 2: Fit D regressions with the model M_t on D imputed datasets. Calculate the combine *p-values* for variables other than X_a. Denote X_b be the variable with the largest combined *p-value* p_b. If $p_b > \alpha_2$, X_b is dropped from the model M_t.

Step 3: Iterate between *Step 1* and *Step 2* until the procedure terminates.

When the MI-stepwise selection procedures terminates, the combined *p-values* for all the variables in the model are not bigger than α_2 and none of the variables not in the model has combined *p-value* smaller than or equal to α_1 if it is included in the model. The MI-stepwise method yields a consistent selection of important variables via the above iterative procedure. By using Rubin's rules repeatedly in each step of selection, the imputation uncertainty caused by missing data is also incorporated in the variable selection.

4. MI-lasso method

Variable selection based on the penalized likelihood function has become popular in recent statistical literature. In particular, the lasso method (Tibshirani, 1996) has gained much attention in recent years. It penalizes the L_1-norm of the regression coefficients to achieve a sparse model:

$$\min_{\beta_j} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{3}$$

where λ is a non-negative tuning parameter. Due to the singularity at $\beta_j = 0$ of the L_1-norm penalty, some estimated $\hat{\beta}_j$ will be exactly zero, which realizes the variable selection.

When multiply-imputed data are present, if lasso is applied to each imputed dataset separately, it is possible that it selects different variables in different imputed datasets. This inconsistency is due

to fitting the model on each dataset *separately*, and this motivates us to consider fitting models on all imputed datasets *jointly* to yield a consistent variable selection across all imputed datasets.

Denote $\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{D,j}$ be the D estimated coefficients for variable X_j in D imputed datasets. If X_j is unimportant, $\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{D,j}$ should be *all* zero, and if X_j is important, $\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{D,j}$ should be *all* nonzero. Motivated by this desired consistency, we treat $\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{D,j}$ as a group, and apply the group lasso penalty (Yuan and Lin, 2006). To be specific, we consider the following optimization approach:

$$\min_{\beta_{d,j}} \sum_{d=1}^{D} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_{d,j} x_{d,ij})^2 + \lambda \sum_{j=1}^{p} \sqrt{\beta_{1,j}^2 + \cdots + \beta_{D,j}^2}, \tag{4}$$

where $\sum_{j=1}^{p} \sqrt{\beta_{1,j}^2 + \cdots + \beta_{D,j}^2}$ is called the group lasso penalty, and λ is a non-negative tuning parameter. We will talk about how to tune λ in Section 5.

We can see that, due to the group lasso penalty, $\hat{\beta}_{d,j}$, the estimated regression coefficient for X_j in the d^{th} imputed dataset, depends on *all* imputed datasets instead of the d^{th} imputed dataset only. In other words, the proposed method fits D models on all imputed datasets *jointly* instead of fitting models *separately*. Furthermore, owing to the nature of the group lasso penalty (Yuan and Lin, 2006), for each variable X_j, the corresponding estimated coefficients $(\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{D,j})$ will be either *all* exactly zero or *all* nonzero. This property guarantees the desired consistency of variable selection across all imputed datasets. Once the variable selection has completed, Rubin's rules can be applied to produce the overall inference.

It is not trivial to get the solution to the optimization problem (4), because the group lasso penalty function is singular at the origin point. To overcome this optimization difficulty, we apply the local quadratic-approximation method as proposed in Fan and Li (2001). To be specific, we iteratively solve the optimization problem (4). Suppose we already have the estimates $\hat{\beta}_{d,j}^{(t)}$, $d = 1, \ldots, D$, at the t^{th} iteration. As long as $\sqrt{(\hat{\beta}_{1,j}^{(t)})^2 + \cdots + (\hat{\beta}_{D,j}^{(t)})^2} > 0$, we have the following approximation:

$$\sqrt{\beta_{1,j}^2 + \cdots + \beta_{D,j}^2} \approx \frac{\beta_{1,j}^2 + \cdots + \beta_{D,j}^2}{\sqrt{(\hat{\beta}_{1,j}^{(t)})^2 + \cdots + (\hat{\beta}_{D,j}^{(t)})^2}}. \tag{5}$$

Correspondingly, the optimization problem (4) can be approximated by

$$\min_{\beta_{d,j}} \sum_{d=1}^{D} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_{d,j} x_{d,ij})^2 + \lambda \sum_{j=1}^{p} c_j \beta_{d,j}^2 \right\}, \tag{6}$$

where $c_j = 1/\sqrt{(\hat{\beta}_{1,j}^{(t)})^2 + \cdots + (\hat{\beta}_{D,j}^{(t)})^2}$. We can see that in the optimization problem (6), the estimated regression coefficients $\hat{\beta}_{d,j}^{(t+1)}$'s can be obtained by solving D separate ridge regressions, which is easy to implement in practice. The iterations continue until the convergence is claimed. One possible limitation for this approximation is that once a group of coefficients are shrunken to zero, they will stay at zero. In order to avoid this inflexibility, we propose to fix $\hat{\beta}_{1,j}^{(t)} = \cdots = \hat{\beta}_{D,j}^{(t)} = \delta$ when $\sqrt{(\hat{\beta}_{1,j}^{(t)})^2 + \cdots + (\hat{\beta}_{D,j}^{(t)})^2} \leqslant \sqrt{D}\delta$. In our algorithm, we choose $\delta = 10^{-10}$, and our numerical studies show that the algorithm performs well.

5. Simulation Study

In this section, we use simulation studies to demonstrate our MI-stepwise method and MI-lasso method for variable selection with multiply-imputed data.

5.1 *Design of Simulation Study*

In our simulations, there are $n = 100$ subjects and $p = 20$ variables, which are generated from a multivariate normal distribution with zero mean, unit variance, and a compound symmetric correlation structure, i.e., $cov(X_i, X_j) = \rho,\ \forall i \neq j$. We consider both $\rho = 0.1$ for low correlation and $\rho = 0.5$ for high correlation. Among 20 variables, the variables $(X_1, X_2, X_3, X_{11}, X_{12}, X_{13})$ are important, and the true model is

$$Y_i = 0.5X_{i1} + X_{i2} + 2X_{i3} + 0.5X_{i11} + X_{i12} + 2X_{i13} + \epsilon_i,\ i = 1, \ldots, 100, \tag{7}$$

where $\epsilon_i \sim N(0, \sigma^2)$. The value of σ^2 is chosen so that the signal-to-noise ratio (SNR) of the model is 3.

We studied two ignorable missing-data mechanisms, missing completely at random (MCAR) and missing at random (MAR). We also considered two different nonresponse rates, which resulted in

60% and 35% of complete cases. For MCAR, since the missingness does not depend on any data value, we dropped 2.5% and 5% of cases independently from each variable in $X_1, \ldots, X_{20}$ to generate incomplete datasets with 60% and 35% complete cases, respectively. While under MAR, the missingness depends on the observed data but not on the missing components. In our simulation studies, the first ten variables, $X_1, \ldots, X_{10}$, were fully observed, and there were missing values in each of the remaining 10 variables, $X_{11}, \ldots, X_{20}$. The missing indicators among these variables were generated using the following logistic regression model:

$$\text{logit}\{\Pr(R_{ij} = 0 | X_{i(j-10)}, Y_i)\} = \alpha_0 + 0.5X_{i(j-10)} + 0.5Y_i, \tag{8}$$

where $j = 11, \ldots, 20$, and α_0 was chosen to yield 60% and 35% of complete cases, respectively.

For each setup, 500 replicates of simulations were conducted. In each replicate, five imputed datasets were generated by chained equations in each incomplete data using R package "MICE" (Van Buuren and Groothuis-Oudshoorn, 2011). Eight methods were compared. The first four methods are based on the stepwise method: the forward stepwise on the original dataset before generating any missing values (SW.O), the forward stepwise on the complete-cases, i.e., the dataset discarding cases with any missing values (SW.C), the forward stepwise based on the Wood's principle (Wood et al. 2008), i.e., the dataset with all five imputed datasets stacked together (SW.stack), and the MI-stepwise on the multiple imputed datasets (SW.M). All the stepwise methods used significance levels of $\alpha_1 = 0.05$ and $\alpha_2 = 0.06$. The other four methods are based on the lasso method: the lasso on the original data before generating any missing values (lasso.O), the lasso on the complete-cases (lasso.C), the lasso on the stacked dataset (lasso.stack), and the MI-lasso on the multiple imputed datasets (lasso.M). Lasso.O, lasso.C, and lasso.stack were fitted using R package "glmnet" (Friedman, Hastie, and Tibshirani, 2010). For all of the regularization methods, the tuning parameters were selected using BIC. In the stacked dataset, each observation was assigned a weight of $1/D$ (Wood et al. 2008). The performance of both the stepwise and the lasso methods on original datasets were provided to serve as a benchmark for the comparison.

For the lasso method, BIC has the following formula:

$$\text{BIC} = \log\Big(\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\hat{\beta}_j x_{ij})^2/n\Big) + df * \log(n)/n, \tag{9}$$

where df is the degrees of freedom of the fitted model. Zou, Hastie and Tibshirani (2008) proved that df can be estimated by the number of nonzero fitted coefficients. For the MI-lasso method, the corresponding BIC is

$$\text{BIC} = \log\Big(\sum_{d=1}^{D}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\hat{\beta}_{d,j} x_{d,ij})^2/(Dn)\Big) + df * \log(Dn)/(Dn), \tag{10}$$

where df is the degrees of freedom of the fitted model. Following Yuan and Lin (2006), it can be shown that the degrees of freedom of our fitted model can be estimated by

$$df = \sum_{j=1}^{p} \boldsymbol{I}(\|\hat{\boldsymbol{\beta}}_j\|_2 > 0) + \sum_{j=1}^{p} \frac{\|\hat{\boldsymbol{\beta}}_j\|_2}{\|\tilde{\boldsymbol{\beta}}_j\|_2}(D-1), \tag{11}$$

where $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{D,j})^T$ and $\tilde{\boldsymbol{\beta}}_j = (\tilde{\beta}_{1,j}, \ldots, \tilde{\beta}_{D,j})^T$ with $\tilde{\beta}_{d,j}$ be the ordinary least square estimate for the j^{th} variable on the d^{th} dataset. The details of derivations are provided in Appendix.

To compare the performance of variable selection among methods, we consider two criteria: sensitivity of selection and specificity of selection, which are defined as follows:

$$\text{Sensitivity} = \frac{\#\text{ of selected important variables}}{\#\text{ of true important variables}}, \tag{12}$$

$$\text{Specificity} = \frac{\#\text{ of removed unimportant variables}}{\#\text{ of true unimportant variables}}. \tag{13}$$

For both sensitivity and specificity, a higher value means better selection performance.

5.2 *Results of Simulation Study*

Table 1 summarizes the sensitivity and specificity of selection for the eight methods. First, the MI-stepwise yields higher sensitivity than the stepwise method on complete cases, and the improvement becomes more significant as either the correlation among variables increases or the percentage of missingness increases. The MI-stepwise also has slightly higher specificity than the stepwise method on the complete cases. Compared to the stepwise method applied on the original dataset before generating any missing values (served as a benchmark), the MI-stepwise has similar

specificity but lower sensitivity of selection. Second, the MI-lasso has a little bit lower specificity but gains a lot in sensitivity compared to the lasso method on the complete-cases. The MI-lasso has similar or even slightly higher sensitivity but lower specificity compared to the lasso method on the original dataset. Finally, the variable selection methods applied on the stacked datasets yield relatively high sensitivity but very low specificity.

[Table 1 about here.]

To further illustrate the selection performance, Figures 1 and 2 show the charts comparing the selection frequency for all important variables ($X_1, X_2, X_3, X_{11}, X_{12}, X_{13}$) and two randomly selected unimportant variables (X_{10}, X_{20}) among different methods under the scenario of MAR. Figure 1 presents the comparison among the four stepwise methods, and Figure 2 presents the comparison among the four lasso methods. The bar charts show that the important variables with larger regression coefficients were more frequently selected into the model than the important variables with smaller regression coefficients. Specifically, X_3 and X_{13} have higher percentages of selection than X_2 and X_{12}, and X_2 and X_{12} have higher percentages of selection than X_1 and X_{11}. Similar to the findings in Table 1, both the MI-stepwise and the MI-lasso result in higher percentages of selection for all the six important variables than the stepwise and lasso applied to complete-cases. Although the stacked data approach has the highest percentage of selection for important variables, it often selects much more unimportant variables than other methods. Furthermore, since the variables $X_1, \ldots, X_{10}$ were designed to be complete but missing data were generated in each of the variables $X_{11}, \ldots, X_{20}$, the percentages of selection are different between these two sets of variables. For two important variables with the same coefficient values such as X_1 and X_{11}, we can see that X_1 has a higher selection percentage than X_{11} when the MI-stepwise was used, while the difference in the selection percentages is negligible when the MI-lasso was used. On the other hand, for the two unimportant variables X_{10} and X_{20}, the percentage of correct

exclusion of X_{10} is higher than that in X_{20} when the MI-lasso was used, while this difference was not observed when the MI-stepwise was used.

[Figure 1 about here.]

[Figure 2 about here.]

6. Application

The UMDES sought to identify factors which explain variations in serum dioxin concentrations from candidate factors that reflect potential exposure to dioxins through air, water, soil, food intake, occupations, and various recreational activities. In this application, we study the human population who resided in or near the flood plain of the Tittabawassee between the Dow Chemical plant in Midland and the confluence of the Tittabawassee and Shiawassee Rivers in Saginaw. We focus on serum 2,3,7,8-tetrachlorodibenzop-dioxin (TCDD), which is the most toxic dioxin compound. In total, 448 eligible study participants had serum TCDD measured. A logarithm 10 transformation was taken on serum TCDD concentration. We applied both the MI-stepwise and MI-lasso methods to select important exposure factors which are related to serum TCDD concentration. Significance levels of $\alpha_1 = 0.05$ and $\alpha_2 = 0.06$ were used in the MI-stepwise method, and BIC was used to tune λ in the MI-lasso method.

The variables selected using the MI-stepwise and the MI-lasso are shown side by side in Table 2. For each variable selected into the model, the estimate of the regression coefficient and its 95% confidence interval (CI) were calculated using Rubin's rules. The top part of the table lists the twelve variables which were selected by both methods, and the bottom part shows the variables which were selected by one method but not the other. The twelve variables which were selected by both methods are all statistically significant at the significant level of 0.05, and their estimates of regression coefficients are similar in the two models. The fitted models suggest that serum TCDD concentration is higher among old people and female and lower among the white people, breast fed

women, people smoke, and people who gained weight in the past one year. The above findings on the demographics and health factors are consistent with the findings in other literature (Patterson et al., 2004; Wittsiepe et al., 2000). More importantly, the fitted models also suggest some local TCDD exposure factors. Specificly, the serum TCDD concentration is higher among people living longer in the Midland county in 1960-1979, fishing frequently in Saginaw river and bay after 1980, using wood burning stoves in 1960-79, or working in paper industry after 1980. In addition to the twelve common variables, three other different variables were also selected by the MI-lasso and the MI-stepwise. These variables need further investigation.

[Table 2 about here.]

7. Conclusion

In this paper, we introduce two novel variable selection methods for multiply-imputed data. Both the MI-stepwise and the MI-lasso methods yield a consistent variable selection via *jointly* fitting models across multiple imputed datasets. Our methods also take into account the imputation uncertainty due to missing data. Simulation studies show that both the MI-stepwise and the MI-lasso methods have higher sensitivity of selecting important variables than the stepwise and the lasso methods applied to the complete-cases. Moreover, the two MI methods perform similarly to the corresponding methods applied on the original dataset before generating any missing data. Our investigation suggests against conducting variable selection on complete-cases, which may fail to identify important variables.

In this article, we focus on missing data with an ignorable missing mechanism. The missing data with nonignorable missing mechanism requires more complex imputation methods. Once the missing data are properly imputed, our two MI variable selection methods are still applicable. Both SAS macros and R codes to implement our MI-stepwise and MI-lasso methods are available upon request.

ACKNOWLEDGEMENTS

The authors thank Dr. David Garabrant for making the UMDES data available and helping in interpreting study results. The authors also thank Dr. Jun Shao for very helpful discussion. This research was supported in part by the Calderone Junior Faculty Research Prize from Columbia University Mailman School of Public Health.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19,** 716-723.

Arisawa K., Takeda H., and Mikasa H. (2005). Background exposure to PCDDs/PCDFs/PCBs and its potential health effects: a review of epidemiologic studies. *The Journal of Medical Investigation* **52,** 10-21.

Baccarelli, A., Mocarelli, P., Patterson Jr., D.G., Bonzini, M., Pesatori, A.C., Caporaso, N., and Landi, M.T. (2002). Immunologic Effects of Dioxin: New Results from Seveso and Comparison with Other Studies. *Environmental Health Perspectives* **110,** 1169-1173.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics* **24,** 2350-2383.

Dalton, T., Kerzee, J., Wang, B., Miller, M., Dieter, M., Lorenz, M., Shertzer, H., Nebert, D., and Puga, A. (2001). Dioxin exposure is an environmental risk factor for ischemic heart disease. *Cardiovascular Toxicology* **1,** 285-298.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani R. (2004). Least angle regression. *Annals of Statistics* **32,** 407-499.

Efroymson, M. A. (1960). Multiple regression analysis. In Ralston, A. and Wilf, HS, editors, *Mathematical Methods for Digital Computers.* New York: Wiley.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96,** 1348-1360.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33,** 1-22.

Garabrant, D.H., Fransblau, A.,Lepkowski J., Gillespie, B.W., Adriaens, P., Demond, A., Ward, B., Ladronka, K., Hedgeman, E., Knutson, K., Zwica, L., Olson, K., Towey, T., Chen, Q., and Hong, B. (2009). The University of Michigan Dioxin Exposure Study: methods for an environmental exposure study of polychlorinated dioxins, furans, and biphenyls. *Environmental Health Perspectives* **117,** 803-810.

Garabrant, D.H., Fransblau, A., Lepkowski, J., Gillespie, B.W., Adriaens, P., Demond, A., Hedgeman, E., Knutson, K., Zwica, L., Olson, K., Towey, T., Chen, Q., Hong, B., Chang, C.W., Lee, S.Y., Ward, B., Ladronka, K., Luksemburg , W., and Maier, M. (2009). The University of Michigan Dioxin Exposure Study: predictors of human serum dioxin concentrations in Midland and Saginaw, Michigan. *Environmental Health Perspectives* **117,** 818-824.

George, E. and Foster, D. (2000). Calibration and Empirical Bayes Variable Selection. *Biometrika* **87,** 731-747.

George, E. and McCulloch, R. (1993). Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association* **88,** 881-889.

Hedgeman, E., Chen, Q., Hong, B., Chang, C.W., Olson, K., Ladronka, K., Ward, B., Adriaens, P., Demond, A., Gillespie, B.W., Lepkowski, J., Franzblau, A., and Garabrant, D.H. (2009). The University of Michigan Dioxin Exposure Study: population survey results and serum concentrations for Polychlorinated Dioxins, Furans, and Biphenyls. *Environmental Health Perspectives* **117,** 811-817.

Heymans, M.W., Van Buuren, S., Knol, D.L, Van Mechelenen, W., and de Vet, H.C.W. (2007).

Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Medical Research Methodology* **7,** 33-42.

Kutner, M., Nachtsheim, C., Neter, J., and Li, W. (2004). *Applied Linear Statistical Models, 5th edition.* McGraw-Hill/Irwin.

Little, J.A. and Rubin, D.B. (2002). *Statistical analysis with missing data, 2nd edition.* Wiley.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics* **37,** 3498-3528.

Meng, X.L. (1995) Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science* **10,** 538-573.

Olson, K., Sinibaldi, J., Lepkowski, J. M, Ward, B., Ladronka, K, Towey, T, Wright, D., and Gillespie, B.W. (2006) Missing data in an environmental exposure study: imputation to improve survey estimation. *Organohalogen Compounds* **68,** 1346-1349.

Park, T. and Casella G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103,** 681-687.

Passi, S., Nazzaro-Porro, M., Boniforti, L., and Gianottij, F. (1981). Analysis of lipids and dioxin in chloracne due to tetrachloro-2,5,7,8-p-dibenzodioxin. *British Journal of Dermatology* **105** 137-143.

Patterson Jr., D.G, Patterson, D., Canady, R., Wong, L.-Y., Lee, R., Turner, W., Caudil, S., Needham, L., and Henderson, A. (2004) Age specific dioxin TEQ reference range *Organohalogen Compounds* **66,** 2844-2849.

Raghunathan T.E., Lepkowski J.M., Van Hoewyk J., 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. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin D.B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of American Statistical Association* **81,** 366-374.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data.* Chapman & Hall, London.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics* **6,** 461-464.

Shen, X. and Huang, H. (2010). Grouping pursuit in regression. *Journal of American Statistical Association*, To appear.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B* **58,** 267-288.

Van Buuren, S., and Groothuis-Oudshoorn, K. (2011) MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, forthcoming. http://www.stefvanbuuren.nl/publications/MICEinR-Draft.pdf

Wittsiepe J, Schrey P, Ewers U, Selenka F, and Wilhelm M. (2000). Decrease of PCDD/F levels in human blood from Germany over the past ten years (1989-1998). *Chemosphere* **40,** 1103-1109.

Wood A.M., White I.R., and Royston P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine* **27,** 3227-3246.

Yang X., Belin T.R., and Boscardin W.J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics* **61,** 498-506.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68,** 49-67.

Zambon, P., Ricci, P., Bovo, E., Casula, A.,Gattolin,M., Fiore, A. R., Chiosi, F., and Guzzinati, S.

(2007). Sarcoma risk and dioxin emissions from incinerators and industrial plants: a population-based case-control study (Italy). *Environmental Health: A Global Access Science Source* **6,** 19.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38,** 894-942.

Zhao, P., Rocha, G. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics* **37,** 3468-3497.

Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B* **67,** 301-320.

APPENDIX

Degrees of freedom estimation for MI-lasso

In this Appendix, we re-represent our MI-lasso method as a group lasso regression problem and derive its degrees of freedom for the fitted model.

For the i^{th} subject, let $\boldsymbol{z}_i$ be a D-length vector with all elements be y_i, i.e., $\boldsymbol{z}_i = (y_i, \ldots, y_i)^T$. Denote $\boldsymbol{u}_{d,ij}$ be a D-length vector with only the d^{th} element be $x_{d,ij}$ and all other elements be zero, i.e., $\boldsymbol{u}_{d,ij} = (0, \ldots, x_{d,ij}, \ldots, 0)^T$. Denote $\boldsymbol{w}_i = (\boldsymbol{u}_{1,ij}, \ldots, \boldsymbol{u}_{D,ij})$. Denote $\boldsymbol{\beta}_j = (\beta_{1,j}, \ldots, \beta_{D,j})^T$. Given λ, the optimization problem (4) is equivalent to the following group lasso regression estimation:

$$\min_{\boldsymbol{\beta}_j} \sum_{i=1}^{n} \|\boldsymbol{z}_i - \sum_{j=1}^{p} \boldsymbol{w}_i^T \boldsymbol{\beta}_j\|_2^2 + \lambda \sum_{j=1}^{p} \|\boldsymbol{\beta}_j\|_2 \tag{A.1}$$

This equivalence can be verified using matrix calculation straightforwardly, and the details are omitted here.

Following the degrees of freedom for group lasso regression which is derived in Yuan and Lin (2006) (equation 6.3), we estimate the degrees of freedom for the fitted model by the MI-lasso as

$$df = \sum_{j=1}^{p} \boldsymbol{I}(\|\hat{\boldsymbol{\beta}}_j\|_2 > 0) + \sum_{j=1}^{p} \frac{\|\hat{\boldsymbol{\beta}}_j\|_2}{\|\tilde{\boldsymbol{\beta}}_j\|_2}(D-1), \tag{A.2}$$

where $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{1,j}, \ldots, \hat{\beta}_{D,j})^T$ and $\tilde{\boldsymbol{\beta}}_j = (\tilde{\beta}_{1,j}, \ldots, \tilde{\beta}_{D,j})^T$ with $\tilde{\beta}_{d,j}$ as the OLS estimate for the j^{th} variable on the d^{th} dataset.

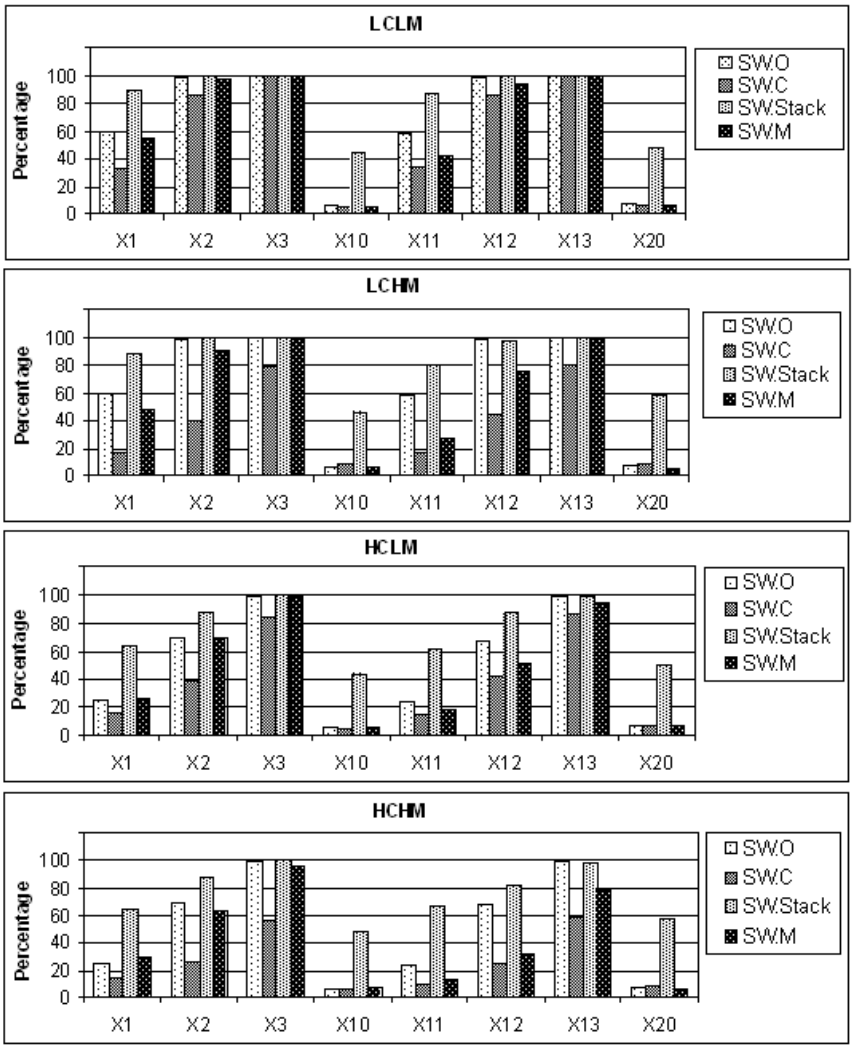

Figure 1. Percentage of simulations in which each variable is selected into the model using the *stepwise selection methods* among the six important variables $(X_1, X_2, X_3, X_{11}, X_{12}, X_{13})$ and two randomly selected unimportant variables (X_{10}, X_{20}). The true model is $Y_i = 0.5X_{i1} + X_{i2} + 2X_{i3} + 0.5X_{i11} + X_{i12} + 2X_{i13} + \epsilon_i$. The first four variables (X_1, X_2, X_3, X_{10}) are fully observed. The last four variables $(X_{11}, X_{12}, X_{13}, X_{20})$ are partially observed, and the missing data mechanisms are MAR. "LCLM": $\rho = 0.1$ and 60% complete cases; HCLM: $\rho = 0.5$ and 60% complete cases; LCHM: $\rho = 0.1$ and 35% complete cases; HCHM: $\rho = 0.5$ and 35% complete cases.

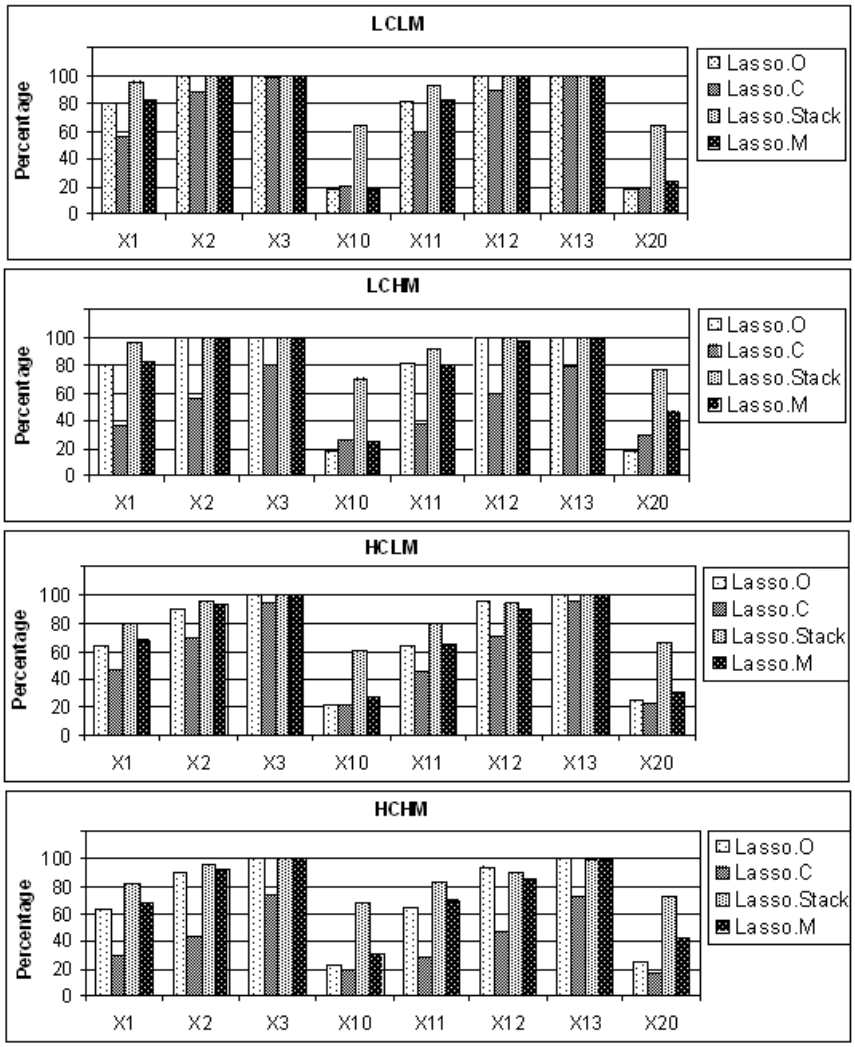

Figure 2. Percentage of simulations in which each variable is selected into the model using the *lasso methods* among the six important variables $(X_1, X_2, X_3, X_{11}, X_{12}, X_{13})$ and two randomly selected unimportant variables (X_{10}, X_{20}). The true model is $Y_i = 0.5X_{i1} + X_{i2} + 2X_{i3} + 0.5X_{i11} + X_{i12} + 2X_{i13} + \epsilon_i$. The first four variables (X_1, X_2, X_3, X_{10}) are fully observed. The last four variables $(X_{11}, X_{12}, X_{13}, X_{20})$ are partially observed, and the missing data mechanisms are MAR. LCLM: $\rho = 0.1$ and 60% complete cases; HCLM: $\rho = 0.5$ and 60% complete cases; LCHM: $\rho = 0.1$ and 35% complete cases; HCHM: $\rho = 0.5$ and 35% complete cases.

Table 1

Average sensitivity and specificity of variable selection in the 500 replicates of simulations . "SEN": sensitivity; "SPE": specificity: "LCLM": $\rho = 0.1$ *and 60% complete cases, "HCLM":* $\rho = 0.5$ *and 60% complete cases, "LCHM":* $\rho = 0.1$ *and 35% complete cases, "HCHM":* $\rho = 0.5$ *and 35% complete cases.*

		MCAR				MAR			
		LCLM	HCLM	LCHM	HCHM	LCLM	HCLM	LCHM	HCHM
SW.O	SEN	85.6	85.6	64.0	64.0	85.6	85.6	64.0	64.0
	SPE	93.7	93.7	93.6	93.6	93.7	93.7	93.6	93.6
SW.C	SEN	74.5	60.9	51.9	38.7	73.0	46.1	47.5	31.6
	SPE	93.6	93.1	93.0	92.4	93.7	92.8	92.9	92.7
SW.stack	SEN	95.7	95.7	84.8	84.0	95.9	94.1	83.8	83.3
	SPE	54.8	53.4	55.1	54.1	53.2	46.3	52.1	47.0
SW.M	SEN	84.1	82.0	63.9	62.3	81.3	73.4	59.5	52.0
	SPE	93.8	94.0	93.8	93.6	94.4	94.6	93.4	92.9
lasso.O	SEN	93.5	93.5	85.6	85.6	93.5	93.5	85.6	85.6
	SPE	82.7	82.7	76.3	76.3	82.7	82.7	76.3	76.3
lasso.C	SEN	86.6	81.2	76.9	68.7	82.0	58.1	70.1	48.9
	SEP	80.5	69.8	76.2	70.3	80.4	72.9	78.2	81.4
lasso.stack	SEN	98.6	98.4	91.7	92.2	98.1	97.8	91.3	91.9
	SEP	36.2	34.1	42.8	38.8	34.4	25.8	37.0	29.6
lasso.M	SEN	94.4	94.4	87.8	87.8	94.0	93.2	86.1	85.9
	SEP	80.7	78.7	72.4	71.0	77.6	63.9	69.3	60.9

Table 2

Comparison of the MI-stepwise and the MI-lasso methods in selecting important variables of log_{10} serum TCDD concentration in the UMDES flood plain and near flood plain sample

MI-Stepwise			MI-Lasso		
Var	Estimate	95% CI	Var	Estimate	95% CI
Intercept	0.359	(0.175, 0.543)	Intercept	0.375	(0.180, 0.570)
age	0.014	(0.011, 0.017)	age	0.013	(0.010, 0.017)
female	0.158	(0.087, 0.229)	female	0.145	(0.076, 0.214)
age × female	0.006	(0.002, 0.009)	age × female	0.006	(0.002, 0.010)
months of breast feeding	-0.006	(-0.009, -0.004)	months of breast feeding	-0.006	(-0.008, -0.003)
lifetime pack-yrs smoking	-0.002	(-0.004, -0.001)	lifetime pack-yrs smoking	-0.002	(-0.004, 0.000)
current smoker	-0.156	(-0.265, -0.048)	current smoker	-0.141	(-0.256, -0.027)
BMI change	-0.020	(-0.029, -0.012)	BMI change	-0.018	(-0.027, -0.009)
yrs of living in Midland: 1960-79	0.009	(0.005, 0.013)	yrs of living in Midland: 1960-79	0.006	(0.000, 0.013)
fishing in Saginaw river and bay after 1980 ($\geqslant$ 1/moth)	0.171	(0.052, 0.290)	fishing in Saginaw river and bay after 1980 ($\geqslant$ 1/month)	0.194	(0.073, 0.314)
white	-0.226	(-0.418, -0.035)	white	-0.218	(-0.421, -0.014)
yrs of using wood burning stoves in 1960-1979	0.008	(0.002, 0.013)	yrs of using wood burning stoves in 1960-1979	0.009	(0.003, 0.014)
yrs of working in paper industry after 1980	0.009	(0.004, 0.013)	yrs of working in paper industry after 1980	0.007	(0.000, 0.014)
BMI	0.006	(0.002,0.011)	ever smoke	-0.050	(-0.128, 0.028)
yrs of spraying chemicals to kill plants after 1980	0.009	(0.002, 0.015)	air dioxin concentration in 1940-59	1.352	(-0.775, 3.480)
yrs of living in a property with crops livestock or poultry after 1980	0.010	(0.003,0.017)	air dioxin concentration in 1960-83	0.521	(-0.861, 1.902)