Published in final edited form as:

Wiley Interdiscip Rev Comput Stat. 2017 ; 9(5): . doi:10.1002/wics.1402.

Variable Selection in the Presence of Missing Data: Imputation-based Methods

Yize Zhao* and Qi Long†

*Department of Healthcare Policy and Research, Weill Cornell Medical College, Cornell University

†Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania

Abstract

Variable selection plays an essential role in regression analysis as it identifies important variables that associated with outcomes and is known to improve predictive accuracy of resulting models. Variable selection methods have been widely investigated for fully observed data. However, in the presence of missing data, methods for variable selection need to be carefully designed to account for missing data mechanisms and statistical techniques used for handling missing data. Since imputation is arguably the most popular method for handling missing data due to its ease of use, statistical methods for variable selection that are combined with imputation are of particular interest. These methods, valid used under the assumptions of missing at random (MAR) and missing completely at random (MCAR), largely fall into three general strategies. The first strategy applies existing variable selection methods to each imputed dataset and then combine variable selection results across all imputed datasets. The second strategy applies existing variable selection methods to stacked imputed datasets. The third variable selection strategy combines resampling techniques such as bootstrap with imputation. Despite recent advances, this area remains under-developed and offers fertile ground for further research.

Keywords

MAR; MCAR; MNAR; imputation; bootstrap; variable selection; missing data; resampling

Introduction

Missing data are commonly encountered in many studies and complicates data analyses. It is well-known that simple, naive approaches such as complete-case analysis that uses only cases with fully-observed data often lead to biased or less precise results. There is a large body of literature on statistical methods for handling missing data¹⁻³.

In particular, missing data present new challenges for variable selection in regression analysis. Variable selection plays an essential role in regression analysis as it identifies important variables that associated with outcomes and has been shown to improve predictive accuracy of the resulting models. Variable selection methods have been widely investigated for fully observed data including classical model selection techniques⁴, penalization techniques⁵⁻⁷, and Bayesian variable selection techniques⁸⁻¹¹. However, in the presence of

missing data, methods for variable selection need to be carefully designed to account for missing data mechanisms and statistical techniques used for handling missing data. Specifically, there are three common missing data mechanisms, missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR)¹. A naive approach of conducting variable selection using only complete cases has been shown to lead to spurious results.¹² Vast majority of existing methods for variable selection in the presence of missing data are developed under the assumption of MAR and, roughly speaking, can be divided into three groups based on statistical techniques used for handling missing data.

The first group seeks to combine variable selection techniques with inverse probability weighting (IPW) for handling missing data or its extensions including augmented IPW (see Tsiatis, 2006), such as the penalized estimating function approach by Johnson¹³ and Johnson et al.¹⁴ and the EEBoost approach by Wolfson¹⁵. The second group seeks to combine variable selection techniques with likelihood-based methods for handling missing data (see section 8 in Little and Rubin, 2002). Methods belonging to this group typically incorporate variable selection techniques such as lasso⁵ and SCAD⁷ into the expectation-maximization algorithm to maximize the penalized observed data likelihood^{16,17}.

The third group seeks to combine variable selection techniques with imputation methods for handling missing data (see sections 4 and 5 in Little and Rubin, 2002). Imputation is arguably the most widely used approach for handling missing data, due to its ease of use as imputation can be conducted easily using existing software packages.^{18–20} One particularly attractive feature of imputation is that imputing missing values is separate from the subsequent analysis of imputed datasets. As such, it is straightforward to apply an existing variable selection method to each imputed dataset. Though combining imputation and variable selection methods provides a promising direction, there are a number of challenges which have been tackled in the literature. One particular challenge is how to combine the variable selection results across multiple imputed datasets in a principled way to obtain the final variable selection result.

The remainder of the article is focused on reviewing existing variable selection methods that are combined with imputation methods for handling missing data. These methods, valid used under the assumptions of MAR and MCAR, largely fall into three general strategies. The first strategy applies existing variable selection methods to each imputed dataset and then combine variable selection results across all imputed datasets. The second strategy applies existing variable selection methods to stacked imputed datasets. The third variable selection strategy combines resampling techniques such as bootstrap with imputation.

Preliminaries

Problem of Interest

To fix ideas and without loss of generality, we consider a linear regression model with n observations and p predictors

$$\mathbf{y} = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ is the outcome of interest, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p)$ ($n \times p$) is the n by p design matrix with p predictors where p is allowed to be larger than n , $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is the set of regression coefficients, and $\boldsymbol{\epsilon}$ is white noise. Denote by $\mathbf{Z} = (\mathbf{y}, \mathbf{X})$ the complete data and by δ_{ij} the missing data indicator matrix such that δ_{ij} , the (i, j) th element of δ , indicates whether the (i, j) th element of $\mathbf{Z}$ is missing. We further denote the observed components of $\mathbf{Z}$ by $\mathbf{Z}_{obs}$ and the missing components of $\mathbf{Z}$ by $\mathbf{Z}_{mis}$. [The problem of interest is to conduct variable selection for (1) in the presence of missing data in $\mathbf{Z}$, where p may be smaller or greater than n .]

Missing Data Mechanism

The missing data mechanism can be defined by the probability $P(\cdot | \mathbf{Z}, \boldsymbol{\xi})$ where $\boldsymbol{\xi}$ are a set of unknown parameters. MCAR is defined as the missingness of data is independent of all data, i.e. $P(\cdot | \mathbf{Z}, \boldsymbol{\xi}) = P(\cdot | \boldsymbol{\xi})$. MAR is defined as the missingness of data depends on only observed data, i.e. $P(\cdot | \mathbf{Z}, \boldsymbol{\xi}) = P(\cdot | \mathbf{Z}_{obs}, \boldsymbol{\xi})$. MNAR is defined as the missingness of data may depend on both observed and missing components of $\mathbf{Z}$. While naive methods such as complete-case and available-case analysis are appropriate under MCAR, MCAR may be unrealistic in practice. MAR is more realistic than MCAR in practice and there is a large body of literature on handling missing data under MAR. While MNAR is most realistic, it is considerably more challenging to handle than the other two mechanisms. Sensitivity analyses are often used under MNAR. Readers are referred to Little and Rubin¹ for more detailed information regarding missing data mechanisms.]

Imputation

Most of the existing imputation methods require an assumption of MCAR or MAR. The key idea of multiple imputation is to replace each missing value with a set of “plausible” values using its (posterior) predictive distribution conditional on the observed data. It typically entails a sequential two-step procedure with first generating a random draw of $\boldsymbol{\theta}$, the unknown parameter associated with the distribution of $\mathbf{Z}$, from its posterior distribution $P(\boldsymbol{\theta} | \mathbf{Z}_{obs})$, and then imputing $\mathbf{Z}_{mis}$ via a random draw from the conditional predictive distribution. Denote the l -th imputed dataset by $\mathbf{Z}^{(l)} = (\mathbf{Z}_{obs}, \mathbf{Z}_{mis}^{(l)})$ with $l = 1, \dots, L$, where $\mathbf{Z}_{mis}^{(l)}$ denotes the imputed values for the missing component $\mathbf{Z}_{mis}$. Each imputed dataset can be analyzed separately and the final analysis results can be combined across L imputed datasets using Rubin’s Rule (RR), which properly accounts for the uncertainty of imputation.²¹ Of note, besides conducting imputation via probabilistic models,^{18–20} there are also other imputation techniques including Mean-and Mode, K-Nearest Neighbor, Hot-Deck, and Auto-associative Neural Network techniques, etc.²² Of note, the methods reviewed in this article are not restricted to any particular imputation technique.]

Variable selection based on Rubin’s rules

One straightforward approach to conduct variable selection in the presence of missing data is to use Rubin’s rules (RR) repeatedly. As discussed in Wood et al.²³, after obtaining multiple imputed datasets, one can use RR to combine the estimates from each imputed

dataset denoted by $\hat{\theta}_k, k=1, \dots, L$ into an overall estimate as $\tilde{\theta} = \frac{1}{L} \sum_{k=1}^L \hat{\theta}_k$ with the estimated variance $\text{Var}(\tilde{\theta}) = \frac{1}{L} \sum_{k=1}^L V_k + \frac{L+1}{L(L-1)} \sum_{k=1}^L (\hat{\theta}_k - \tilde{\theta})(\hat{\theta}_k - \tilde{\theta})'$, and then use classical variable selection approaches such as forward, backward or stepwise selection methods based on Wald tests constructed using $\tilde{\theta}$ and $\text{Var}(\tilde{\theta})$. However, besides intensive computation, the use of classical variable selection methods based on RR also suffers from other widely known limitations including difficulties in dealing with collinearity, misspecified F distribution, and overfitting. In addition, this approach does not work for high-dimensional data where p is close to or greater than n .]

Variable selection on each imputed dataset

The most straightforward way to combine MI with variable selection method is to implement variable selection on each imputed dataset $\mathbf{Z}^{(l)}$ as if it were performed on complete data, resulting in a set of selected variables denoted by $S^{(l)} \in \{1, \dots, p\}$ ($l = 1, \dots, L$). Such variable selection can be conducted using forward, backward or stepwise selection methods based on Wald tests, likelihood ratio test, and Akaike or Bayesian information criterion, regularization methods such as lasso and elastic net, or Bayesian variable selection approaches. However, the set of selected variables $S^{(l)}$ likely vary across L imputed datasets, so a direct application of RR may not be appropriate for combining variable selection results and a number of approaches have been proposed for addressing this challenge.

Wood et al.²³ investigated several approaches to perform variable selection with multiple imputed data. One straightforward approach is to conduct variable selection on a randomly chosen imputed dataset, say the first one after MI. Though this approach was shown to achieve greater efficiency than directly conducting variable selection on complete cases, it does not account for uncertainty of imputation and may lead to large bias in parameter estimation. A better strategy proposed in Wood et al.²³ is to perform model selection on each imputed dataset separately and define the final set of selected variables as

$S = \left\{ j, \sum_{l=1}^L I(j \in S^{(l)})/L \geq \pi \right\}$, where π is a threshold for selecting a predictor and can be set to say, 0.5. In other words, S includes variables that are selected in πL of the L imputed datasets. Given the active set S , a standard MI procedure is performed with the final result combined by RR for a regression model including only variables in S . This approach was shown to yield more robust results in variable selection and coefficient estimation, but variable selection results can be sensitive to the threshold value π . In addition, Wood et al.²³ investigated only classical stepwise model selection based on Wald-test, [which also suffers from the limitations mentioned in the discussion of variable selection based on Rubin's rules.] Their strategies do not work for high-dimensional data with $n < p$. Similar strategies have also been adopted for analysis of longitudinal data in the presence of missing data, where generalized estimating equations (GEE)²⁴ for longitudinal data analysis are used. Shen and Chen²⁵ focused on the model selection criteria by extending the quasi-likelihood under the independence model criterion (QIC)²⁶ and the missing longitudinal information criterion (MLIC)²⁷, to the context of MI when a specific model was applied across L MI

datasets. Specifically, they proposed to use the average QIC or MLIC values across all L imputed datasets to conduct variable selection.

Similar to the strategy assessed in Wood et al.²³, Lachenbruch²⁸ investigated an approach that used least angle regression (LARS)²⁹ and lasso to conduct variable selection in each imputed dataset and then combined the variable selection results using a pre-specified threshold π . However, Lachenbruch²⁸ assessed this approach along with the strategy in Wood et al.²³ only in a data application, but not in simulations.

An alternative strategy is to combine multiple imputation and variable selection in a Bayesian framework. Bayesian variable selection methods automatically allow for quantification of variable selection uncertainty, which is an appealing feature. In addition, since multiple imputation is inherently Bayesian, it seems natural to combine MI and variable selection in a Bayesian framework, though there has been fairly limited work in this direction. Yang et al.³⁰ proposed two approaches to conduct imputation and Bayesian variable selection under linear regression model with missing covariates. The first approach is known as “impute, then select (ITS). Specifically, for each imputed dataset $\mathbf{Z}^{(l)}$, Bayesian Stochastic Search Variable Selection (SSVS)^{8,9} is performed using a latent selection

indicator $\gamma_j^{(l)}$ for $x_j, j = 1, \dots, p$ where $\gamma_j^{(l)} = 1$ if variable x_j is selected and $\gamma_j^{(l)} = 0$ otherwise. Using a Markov Chain Monte Carlo algorithm, one can obtain the posterior probability of $\gamma_j^{(l)} = 1$ denoted by $p_j^{(l)}$. The final variable selection is determined by whether the average

marginal posterior probability of γ across L MI datasets defined as $p_j = \sum_{l=1}^L p_j^{(l)} / L$ exceeds a certain threshold π . This approach allows for quantification of model selection uncertainty using RR. Jiang et al.³¹ further developed an extension to the ITS approach for performing prediction on MI datasets, where prediction can be readily made using posterior distributions. Since MI is inherent Bayesian, Yang et al.³⁰ proposed a second strategy that embeds steps of imputation and SSVS in a single combined Gibbs sampling process, known as Simultaneously Impute and Select” (SIAS). One advantage of SIAS is that it enables quantification of variable selection uncertainty naturally using the posterior distribution of γ_j , the selection indicator for x_j , without resorting to RR. In other words, the final result on variable selection can be directly determined by the marginal posterior probability of the selection indicator, avoiding inconsistency of selection across different MI datasets. Yang et al.³⁰ demonstrated in simulation studies that SIAS slightly outperforms ITS and provides smaller standard errors. However, it is easier to implement ITS using existing commercial software packages and ITS also has the flexibility of having the imputation step and selection step done by different analysts at different times.

Variable selection on stacked imputed dataset

An alternative way to incorporate variable selection with MI is to stack the L imputed datasets to form one large dataset denoted by $\mathbf{Z} = ((\mathbf{Z}^{(1)})^T, \dots, (\mathbf{Z}^{(L)})^T)^T$. Given $\mathbf{Z}$, one can use one of the existing variable selection method. One of the challenges associated with this approach is that though the estimates based on the stacked dataset $\mathbf{Z}$ are consistent, their standard errors are under-estimated due to the artificially inflated sample size. Therefore,

further adjustments are needed to correct standard error estimates. Wood et al.²³ proposed to apply a fixed weight w to all individuals where the sum of all weights would be equal to n or $(1 - f)n$ with f the fraction of missing data—and several possible choices of weights were considered. Then classical backward variable selection was conducted for the weighted regression model. More recently, Wan et al.³² adopted this weighted strategy on the stacked imputed dataset and proposed to use penalization, namely, elastic net³³, to conduct variable selection. Specifically, similar to Wood et al.²³, $\mathbf{Z}$ was created after imputation and a

subject-specific weight $w_i = \frac{f_i}{L}$ is assigned to subject i . The penalized likelihood function for the weighted regression model using the elastic net penalty is used to obtain the parameter estimates.

Chen and Wang¹² took a different perspective when conducting variable selection using the stacked imputed dataset. They considered fitting model (1) in all L imputed dataset $\mathbf{Z}^{(l)}$ jointly by using the joint likelihood, i.e., the product of the likelihood functions across all L imputed datasets where $\beta^{(l)}$ denotes the regression coefficients in model (1) in $\mathbf{Z}^{(l)}$. Their goal is to ensure variable selection consistency across all imputed datasets, i.e., the active set $S^{(l)}$ in each imputed dataset $\mathbf{Z}^{(l)}$ is the same. At the same time, their approach allows the estimates of nonzero elements of $\beta^{(l)}$ to be different. To achieve their goal, they treated the

estimates of each coefficient β_j in all L imputed datasets denoted by $\hat{\beta}_j^{(1)}, \dots, \hat{\beta}_j^{(L)}$ as a group, and apply the group lasso penalty³⁴ to each group for all p variables in the joint

likelihood. Due to the grouping nature of the group lasso penalty, $\hat{\beta}_j$ obtained from the L imputed datasets should all be zero if X_j is important or nonzero otherwise, which guarantees the desired consistency of variable selection across all imputed datasets.

Geronimi and Saporta³⁵ further extended the approach by Chen and Wang¹² to the analysis of longitudinal data in which GEE was used. As such, the joint penalized likelihood function in Chen and Wang¹² was replaced by the joint penalized generalized estimating equations with the group lasso penalty where the coefficients for the same predictor x_j across L imputed datasets were considered as a group.

Variable selection on imputed dataset combined with resampling

Resampling techniques such as bootstrap³⁶ are widely used to quantify uncertainty in parameter estimation. There are interesting similarities between such resampling techniques and multiple imputation. Both account for key variations by generating multiple datasets, analyzing each dataset separately, and combining the analysis results from individual datasets into a final analysis result. [In addition, in the variable selection literature, it has been shown that various versions of randomized lasso^{37,38} that apply lasso to resampled datasets using bootstrap³⁶ or random subsampling³⁷ has several advantages over the standard lasso.] In particular, compared to the standard lasso, variable selection results using randomized lasso are less sensitive to the amount of regularization and is consistent in variable selection under a condition on the design matrix that is weaker than the irreducible condition needed for the model selection consistency of lasso³⁹. As a result, there is a growing body of literature on combining resampling techniques with imputation for variable selection in the presence of missing data.

Heymans et al.⁴⁰ investigated how MI and bootstrapping can be properly combined into the selection process of prognostic variables in the presence of missing data and whether this influences model performance through a case study. They explored a number of options for combining MI and bootstrapping in their numerical studies in terms of the influence of resampling and imputation variation. Among the approaches assessed, they focused specifically on first conducting multiple imputation and then constructed bootstrap datasets by randomly drawing with replacement from each imputed dataset. Then a classical backward selection based on Wald-test was applied to each bootstrap dataset. To determine the final set of selected variables, they adopted an approach similar to Wood et al.²³, i.e., using a threshold on the inclusion frequency of each variable across all bootstrap datasets. In addition, they also considered other competing methods including using MI alone and constructing bootstrap datasets on the first imputed dataset only, which shows less promising results in their numerical studies compared with the combining approach. However, their comparisons of the methods that were considered are inconclusive due to a lack of robust simulation studies.

The obvious drawbacks of the methods in Heymans et al.⁴⁰ are the use of classical variable selection which suffers from the drawbacks as discussed previously, and from the fact that variable selection results are sensitive to the threshold value. A more recent work by Musoro et al.⁴¹ sought to improve the methods in Heymans et al.⁴⁰ by using penalization, namely, lasso, for variable selection on each bootstrap dataset. In their work, they also focused on a case study of a cohort of Chronic Obstructive Pulmonary Disease patients, and considered and compared different approaches on how to combine MI and bootstrap before conducting variable selection via lasso using each individual dataset. Specifically, they considered three variations of the approach that generates bootstrap samples from each imputed dataset $\mathbf{Z}^I$ and a fourth approach of generating bootstrap samples from the incomplete data $\mathbf{Z}$ followed by MI. Of note, in their work, bootstrap was used primarily for quantifying optimism⁴². Using the same variable selection technique, they compared the four approaches in their numerical study in terms of the optimism value and mean squared error, and the results are mixed with the fourth approach generally achieving better performance. Hence such approach is recommended to deal with prognostic models though there are no theoretical justifications.

To take advantage of randomized lasso, a recent work by Long and Johnson⁴³ investigated variable selection in the presence of missing data by combining resampling techniques with imputation. Their approach, known as BI-SS, starts by generating bootstrap datasets $\mathbf{Z}_1, \dots, \mathbf{Z}_B$ from the original data $\mathbf{Z}$ and then conduct imputation on each bootstrap dataset, which mimics bootstrap imputation proposed by Efron⁴⁴. For each bootstrap imputed dataset, regression coefficients β are estimated via randomized lasso. The final set of selected variables is determined by stability selection³⁷ using a pre-specified threshold value π . In their work, they also investigated an alternative approach by using bootstrapped lasso proposed by^{45,46}, which is known as BI-BL. BI-BL entails performing variable selection for each bootstrap imputed dataset $\mathbf{Z}_b$ and then determining the set of selected variables S by intersecting the active sets across all bootstrap imputed dataset. Their numerical studies included general missing data patterns and both low-dimensional ($n > p$) and high-

dimensional ($n < p$) settings and demonstrated that the BI-SS approach achieves better performance in variable selection and prediction than BI-BL as well as several methods proposed in Heymans et al.⁴⁰, Lachenbruch²⁸ and Johnson et al.¹⁴.

Similar in spirit to Long and Johnson⁴³, Liu et al.⁴⁷ proposed a multiple imputation random lasso (MIRL) method. The key difference between their approach and the BI-SS approach in Long and Johnson⁴³ is that their approach first generates multiply imputed datasets $\mathbf{Z}^{(l)}$.

Specifically, under each $\mathbf{Z}^{(l)}$, B bootstrap samples are generated denoted by $\{\mathbf{Z}_1^{(l)}, \dots, \mathbf{Z}_B^{(l)}\}$,

and lasso-OLS estimates $\hat{\beta}_b^{(l)}$ are obtained from each $\mathbf{Z}_b^{(l)}$ with $l = 1, \dots, L$; $b = 1, \dots, B$ in order to calculate an overall importance measure for each variable. Half of the variables are then randomly selected based on importance measure with coefficients computed by lasso-OLS and the initial MIRL estimators is set to be the average of the estimates across bootstrap samples and imputations. Finally, the variable selection will be refined by stability selection determined by the ranks of importance measure. The choice of random lasso³⁸ is due to the claimed advantage for dealing with highly correlated predictors. Their numerical studies compared the MIRL method with the methods in Johnson et al.¹⁴, Chen and Wang¹², and Garcia et al.¹⁷, and demonstrated that the MIRL method achieves better performance in the presence of a high proportion of missing data and highly correlated predictors.

One strength of the BI-SS⁴³ and MIRL⁴⁷ methods is that they can handle high-dimensional data settings ($p > n$). However, numerical studies directly comparing these two methods have not been reported in the literature.

Discussion

This article has focused on reviewing statistical methods for variable selection in the presence of missing data where imputation is used for handling missing data under the assumption of MAR. One advantage of these methods is that they combine existing imputation methods with existing advanced variable selection methods, making their implementation straightforward. Due to the popularity and wide applicability of imputation methods, these methods are appealing to practitioners. [In particular, the current empirical results suggest that the approaches proposed by Chen and Wang¹², Long and Johnson⁴³, and Liu et al.⁴⁷ are preferred to the other methods, as these methods are applicable to both low-dimensional and high-dimensional problems and have been shown to outperform the other methods in numerical studies.]

However, the research in this area remains fairly limited relative to variable selection for fully observed data and offers fertile ground for further development. For example, there is a lack of comprehensive, systematic assessments of existing methods through extensive numerical studies, which would provide valuable insights on their strengths and weaknesses. In addition, the existing work in this area has not investigated principled approaches for attaching standard errors to parameter estimates that account for uncertainty of variable selection, though Long and Johnson⁴³ briefly described a possible approach extending the work of Chatterjee and Lahiri⁴⁸. There has been little to no work on variable selection in the case of NMAR, which is expected to be highly challenging.

The performance of these methods relies heavily on the choice of imputation methods and careful construction of imputation models, which is highly challenging in the presence of high-dimensional data. Some recent work by Zhao and Long⁴⁹ and Deng et al.⁵⁰ has investigated imputation methods in the presence of high-dimensional data, but methods in this area are largely under-developed and additional research is urgently needed.

Acknowledgments

Funding

This work was supported in part by PCORI award ME-1303-5840, and NIH grants R21NS091630 and P30CA016520. The content is solely the responsibility of the authors and does not necessarily represent the official views of the PCORI and NIH.]

References

1. Little, R., Rubin, D. Statistical Analysis with Missing Data. John Wiley; New York: 2002.
2. Tsiatis, A. Semiparametric theory and missing data. Springer; 2006.
3. Rässler S, Rubin DB, Zell ER. Imputation. Wiley Interdisciplinary Reviews: Computational Statistics. 2013; 5:20.
4. Akaike H. A new look at the statistical model identification. Automatic Control, IEEE Transactions on. 1974; 19:716.
5. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B. 1996; 58:267.
6. Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association. 2006; 101:1418.
7. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348.
8. George E, McCulloch R. Approaches for Bayesian variable selection. Statistica Sinica. 1997; 7:339.
9. George EI, McCulloch RE. Variable selection via Gibbs sampling. Journal of the American Statistical Association. 1993; 88:881.
10. Kyung M, Gill J, Ghosh M, Casella G. Penalized Regression, Standard Errors, and Bayesian Lassos. Bayesian Analysis. 2010; 5(2):369.
11. Hans C. Model uncertainty and variable selection in Bayesian lasso regression. Statistics and Computing. 2010; 20:221.
12. Chen Q, Wang S. Variable selection for multiply-imputed data with application to dioxin exposure study. Statistics in Medicine. 2013; 32:3646. [PubMed: 23526243]
13. Johnson B. Variable selection in semiparametric linear regression with censored data. J R Statist Soc Ser B. 2008; 70:351.
14. Johnson B, Lin D, Zeng D. Penalized estimating functions and variable selection in semiparametric regression models. Journal of the American Statistical Association. 2008; 103:672. [PubMed: 20376193]
15. Wolfson J. EEBoost: A general method for prediction and variable selection based on estimating equations. Journal of the American Statistical Association. 2011; 106
16. Garcia R, Ibrahim J, Zhu H. Variable selection in the Cox regression model with covariates missing at random. Biometrics. 2009; 66:97. [PubMed: 19459831]
17. Garcia R, Ibrahim J, Zhu H. Variable selection for regression models with missing data. Statistica Sinica. 2010; 20:149. [PubMed: 20336190]
18. Buuren S, Groothuis-Oudshoorn K. MICE: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011; 45
19. Raghunathan T, Lepkowski J, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology. 2001; 27:85.

20. Su Y, Gelman A, Hill J, Yajima M. Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. *Journal of Statistical Software*. 2011; 452
21. Rubin, DB. Multiple imputation for nonresponse in surveys. Vol. 81. John Wiley & Sons; 2004.
23. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? *Statistics in Medicine*. 2008; 27:3227. [PubMed: 18203127]
24. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* pp. 1986:13–22.
25. Shen CW, Chen YH. Model selection of generalized estimating equations with multiply imputed longitudinal data. *Biometrical Journal*. 2013; 55:899. [PubMed: 23970494]
26. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics*. 2001; 57:120. [PubMed: 11252586]
27. Shen CW, Chen YH. Model selection for generalized estimating equations accommodating dropout missingness. *Biometrics*. 2012; 68:1046. [PubMed: 22463099]
28. Lachenbruch PA. Variable selection when missing values are present: a case study. *Statistical Methods in Medical Research*. 2011; 20:429. URL <http://smm.sagepub.com/content/20/4/429.abstract>. [PubMed: 20442196]
29. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *The Annals of Statistics*. 2004; 32:407.
30. Yang X, Belin T, Boscardin W. Imputation and variable selection in linear regression models with missing covariates. *Biometrics*. 2005; 61:498. [PubMed: 16011697]
31. Jiang, G., Tam, CH., Luk, AO., Kong, AP., So, WY., Chan, JC., Ma, RC., Fan, X. 2016 IEEE International Conference on. IEEE; 2016. Bioinformatics and Biomedicine (BIBM); p. 727-730.
32. Wan Y, Datta S, Conklin D, Kong M. Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect. *Journal of Statistical Computation and Simulation*. 2015; 85:1902. [PubMed: 26412909]
33. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2005; 67:301.
34. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2006; 68:49.
35. Geronimi J, Saporta G. Variable selection for multiply-imputed data with penalized generalized estimating equations. *Computational Statistics & Data Analysis*. 2017; 110:103.
36. Efron, B., Tibshirani, RJ. An introduction to the bootstrap. CRC press; 1994.
37. Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2010; 72:417. URL <http://dx.doi.org/10.1111/j.1467-9868.2010.00740.x>.
38. Wang S, Nan B, Rosset S, Zhu J. Random lasso. *The Annals of Applied Statistics*. 2011; 5:468. [PubMed: 22997542]
39. Zhao P, Yu B. On model selection consistency of Lasso. *Journal of Machine learning research*. 2006; 7:2541.
40. Heymans MW, van Buuren S, Knol DL, van Mechelen W, de Vet HC. Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Medical Research Methodology*. 2007; 7:33. [PubMed: 17629912]
41. Musoro JZ, Zwinderman AH, Puhan MA, ter Riet G, Geskus RB. Validation of prediction models based on lasso regression with multiply imputed data. *BMC Medical Research Methodology*. 2014; 14:116. [PubMed: 25323009]
42. Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*. 1996; 15:361. [PubMed: 8668867]
43. Long Q, Johnson BA. Variable selection in the presence of missing data: resampling and imputation. *Biostatistics*. 2015; 16:596. [PubMed: 25694614]
44. Efron B. Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*. 1994; 89:463.

45. Bach F. Bolasso: model consistent Lasso estimation through the bootstrap. Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML). 2008
46. Bach F. technical Report. arXiv. 2009 0901.3202.
47. Liu Y, Wang Y, Feng Y, Wall MM, et al. Variable selection and prediction with incomplete high-dimensional data. The Annals of Applied Statistics. 2016; 10:418. [PubMed: 27213023]
48. Chatterjee A, Lahiri SN. Bootstrapping Lasso Estimators. Journal of the American Statistical Association. 2011; 106:608.
49. Zhao Y, Long Q. Multiple imputation in the presence of high-dimensional data. Statistical Methods in Medical Research. 2016; 25:2021. [PubMed: 24275026]
50. Deng Y, Chang C, Ido MS, Long Q. Multiple imputation for general missing data patterns in the presence of high-dimensional data. Scientific Reports. 2016; 6