Title

How to Apply Variable Selection Machine Learning Algorithms With Multiply Imputed Data: A Missing Discussion

Journal

Psychological Methods, 28(2)

ISSN

1082-989X

Authors

Gunn, Heather J
Rezvan, Panteha Hayati
Fernández, M Isabel
et al.

Publication Date

2023-04-01

DOI

10.1037/met0000478

Peer reviewed

HHS Public Access

Author manuscript

Psychol Methods. Author manuscript; available in PMC 2023 July 01.

Published in final edited form as:

Psychol Methods. 2023 April ; 28(2): 452–471. doi:10.1037/met0000478.

How to Apply Variable Selection Machine Learning Algorithms With Multiply Imputed Data: A Missing Discussion

Heather J. Gunn[1], **Panteha Hayati Rezvan**[2], **M. Isabel Fernández**[3], **W. Scott Comulada**[2]

[1]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota, United States

[2]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles

[3]College of Osteopathic Medicine, Nova Southeastern University

Abstract

Psychological researchers often use standard linear regression to identify relevant predictors of an outcome of interest, but challenges emerge with incomplete data and growing numbers of candidate predictors. Regularization methods like the LASSO can reduce the risk of overfitting, increase model interpretability, and improve prediction in future samples; however, handling missing data when using regularization-based variable selection methods is complicated. Using listwise deletion or an ad hoc imputation strategy to deal with missing data when using regularization methods can lead to loss of precision, substantial bias, and a reduction in predictive ability. In this tutorial, we describe three approaches for fitting a LASSO when using multiple imputation to handle missing data and illustrate how to implement these approaches in practice with an applied example. We discuss implications of each approach and describe additional research that would help solidify recommendations for best practices.

Translational Abstract

Standard linear regression is a commonly used model in psychological research that tests the relationships between hypothesized predictors and an outcome of interest; however, the estimated regression coefficients representing such associations are highly variable from sample to sample, making the conclusions less generalizable. Regularization methods like the LASSO reduce the variance of the estimates, increase model interpretability, and improve prediction in future samples. Until recently, regularization methods were primarily applied on data sets without missing values. Missing data are prevalent in psychological research and need to be handled appropriately to avoid substantial bias. Multiple imputation has gained currency as a principled approach to deal with missing data. This tutorial describes three approaches for fitting a LASSO for variable selection when using multiple imputation to handle missing data, highlighting the additional research needed to solidify recommendations for best practices.

Correspondence concerning this article should be addressed to Heather J. Gunn, Department of Quantitative Health Sciences, Mayo Clinic, 205 3rd Avenue Southwest, Harwick 7-37C, Rochester, MN 55905, United States. Gunn.Heather2@mayo.edu.

The authors declare they have no conflicts of interest.

Keywords

LASSO; missing data; multiple imputation; regularization; regression

Machine learning variable selection methods like the least absolute shrinkage and selection operator, familiarly known as LASSO (Tibshirani, 1996), and elastic net (Zou & Hastie, 2005) have not been widely used in psychological research despite appealing properties such as reduced risk of overfitting, increased model interpretability, and improved prediction in future samples compared with the standard linear regression model (McNeish, 2015). A pragmatic consideration inhibiting the use of machine learning variable selection methods is the complication that arises when data contain missing values, a problem prevalent in psychological research. When estimating a standard linear regression model, there are clear guidelines on how to use modern missing data handling techniques like multiple imputation (MI) to reduce bias and increase power (Enders, 2010). However, when using variable selection techniques like the LASSO, there is a lack of accessible guidance on the numerous decisions that must be made to implement the variable selection process in conjunction with modern missing data handling methods. The goal of this article is to provide a tutorial on how to implement the LASSO when using MI to handle missing values, highlighting three approaches that could be readily extended to the elastic net procedure.

Selecting meaningful predictors of an outcome of interest is a challenging statistical problem that psychologists face (Hesterberg et al., 2008). If all hypothesized predictors are included in a simple linear model simultaneously, this can lead to overfitting, inflation of regression coefficient standard errors, and nonparsimonious models. Variable selection methods identify a set of variables that are most associated with or predictive of the outcome of interest. They are primarily used when it is not feasible to include all the relevant predictors and their interactions in the model. Classical variable selection methods such as backward, forward, or stepwise selection (Harrell, 2001) typically provide an interpretable model, where the best model is selected via significance tests or some form of information-based criterion (e.g., Akaike information criterion [AIC] and Bayesian information criterion [BIC]). However, they have been frequently criticized due to their potential for overfitting, inferior predictive ability, and difficulties with handling collinearity (Harrell, 2001).

Another way to improve standard linear regression is to use regularization techniques (also called penalization or shrinkage methods) to constrain or shrink the regression coefficients. Three common machine learning shrinkage methods are ridge regression, LASSO, and elastic net. Ridge regression uses a penalty parameter to shrink the regression coefficients toward zero but does not fix them to zero. LASSO uses a different penalty parameter that also shrinks the estimated regression coefficients toward zero, but unlike ridge, it shrinks some of the estimated coefficients to exactly zero. Thus, LASSO performs variable selection whereas ridge regression does not. Elastic net combines the penalties of both LASSO and ridge regression and is also considered a variable selection procedure because it typically sets some coefficients to zero.

The elastic net improves performance compared with the LASSO in situations where highly correlated variables exist, particularly in large-scale data sets where the number of variables

is greater than the sample size (Zou & Hastie, 2005). The application of elastic net is commonly found in neuroimaging studies (e.g., Cui & Gong, 2018; Gabrieli et al., 2015; Ryali et al., 2012) and in gene expression microarray data analysis (e.g., De Mol et al., 2009; Waldmann et al., 2013). For instance, in gene selection problems, where the sample size is often small due to the high cost of data collection involving human subjects, and there is a strong dependency between genes sharing a common biological pathway, elastic net tends to select (or omit) all the highly correlated genes as a group. In such scenarios, elastic net may be a better choice when the goal is explanation rather than prediction; when the goal is prediction, LASSO may be preferred since it only selects one predictor from a group of highly correlated predictors.

Research in psychology often focuses on explanation over prediction, but the possibility of improving upon predictions from explanatory models has long been recognized (Hagerty & Srinivasan, 1991; Shmueli, 2010). Yarkoni and Westfall (2017) argue that research in psychology would benefit from the use of machine learning models, which often outperforms traditional implementation of statistical models like standard linear regression. Machine learning models should supplement, not replace, currently used models, expanding the type of research questions psychologists can answer. In this tutorial, we focus on the LASSO, which has been used in a variety of behavioral science studies (Ammerman et al., 2018; Comulada et al., 2021; Dumas et al., 2020; Feng et al., 2020; Harris et al., 2020; Hung et al., 2020; Immekus et al., 2019; Smith et al., 2019), scrutinized in simulation studies (Chen & Wang, 2013; Thao & Geskus, 2019), and recommended for wider use in the behavioral sciences (Johnson & Sinharay, 2011; McNeish, 2015). The approaches discussed in this tutorial could readily be applied to the elastic net as well.

Missing data are inevitable in psychological research and can be expected to affect the precision and generalizability of the results, especially if not handled properly. Many previous research projects have relied on listwise deletion (also known as complete case analysis) or ad hoc imputation approaches such as mean substitution. Listwise deletion excludes cases with any incomplete values, which can greatly reduce the sample size, resulting in a loss of precision and statistical power, and gives biased results if individuals with missing observations differ systematically from those with complete observations (Greenland & Finkle, 1995; Horton & Kleinman, 2007; Sterne et al., 2009). Mean substitution overstates the precision of unobserved values, carrying potential to bias point estimates and understate estimates of variability, thereby exaggerating the precision of parameter estimates.

Modern principled methods for handling missing data include maximum likelihood estimation (Arbuckle, 1996; Beale & Little, 1975; Dempster et al., 1977; Enders, 2010), Bayesian estimation (Gelman et al., 2014), inverse probability weighting (Li et al., 2013; Seaman & White, 2013), and MI (Little & Rubin,2019; Rubin, 1987, 1996; Schafer, 1997; van Buuren, 2018). With expanded computing power and accompanying software development, these methods, which overcome the restriction of listwise deletion and account for the uncertainty surrounding missing data, have become more accessible to researchers in social and behavioral disciplines.

MI, in particular, has been increasingly used as a flexible and accessible approach to address missing data (Hayati Rezvan et al., 2015; Mackinnon, 2010), and is now widely recommended by journal reviewers (Little et al., 2012; Ware et al., 2012). MI includes all available data, but more importantly, it accounts for the uncertainty of missing data by imputing missing values multiple times, which results in multiple imputed data sets. The analysis of interest is conducted on each imputed data set separately, and then the estimates obtained from each imputed data set are aggregated via Rubin's rules (Rubin, 1987).

Given the drawbacks of using listwise deletion or ad hoc imputation methods when performing variable selection, several studies have used maximum likelihood estimation (Garcia et al., 2010a, 2010b; Sabbe et al., 2013), Bayesian estimation (Ibrahim et al., 2008; Makalic & Schmidt, 2016; Park & Casella, 2008; Yang et al., 2005), inverse probability weighting (Johnson, 2008; Johnson et al., 2008; Wolfson, 2011), or MI (Zhao & Long, 2017) to address missing data. Still, it is not uncommon for variable selection with incomplete data to be implemented in behavioral science research using listwise deletion (e.g., Comulada et al., 2021; Comulada et al., 2020; Feelders, 1999; Nam et al., 2020), or ad hoc imputation methods such as median or mean substitution (e.g., Masconi et al., 2015; Pelham et al., 2020; Simon et al., 2013; Smith et al., 2019). This is partly due to limitations of commonly used statistical software programs. For example, glmnet, a package in R software (R Core Team, 2018) that estimates LASSOs, is premised on complete case analysis.

There has been particular focus on using MI to address missing data when performing classical variable selection (Austin et al.,2019; Vergouwe et al., 2010; Wood et al., 2008) and machine learning variable selection (Chen & Wang, 2013; Deng et al., 2016; Lachenbruch, 2011; Thao & Geskus, 2019; Wan et al., 2015). A complication of combining MI with variable selection methods, particularly LASSO, is that conducting variable selection on each imputed data set results in different variables selected across the imputed data sets. There are open questions on how to best aggregate variable selection results across multiple imputed data sets to obtain an overall result. This tutorial illustrates three strategies for estimating a LASSO when using MI to handle missing data and discusses how to incorporate cross-validation and training and test sets in the process, which has received little attention in the literature.

The tutorial is organized as follows. First, we describe the foundations for estimating a LASSO when using multiple imputation to handle missing data by introducing an applied setting that we use as a motivating example throughout the tutorial, reviewing the LASSO procedure assuming there is complete data, and reviewing the MI framework for handling missing data. Then, we consider three approaches for fitting a LASSO with multiply imputed data (henceforth referred to as *imputation LASSO approaches*). We investigate: (a) LASSO using a traditional penalty applied to each of the imputed data sets (henceforth described as *the separate approach*); (b) LASSO using a traditional penalty applied to a stacked version of the imputed data sets (henceforth, *the stacked approach*); and (c) a group LASSO applied to the imputed data sets jointly via the MI-LASSO method (Chen & Wang, 2013). We embed the applied example throughout the review of these statistical methods. We supplement these illustrations with a discussion of implications of

the imputation LASSO approaches and consideration of additional research that would help solidify recommendations for best practices.

Foundations for Tutorial Investigation

Applied Example

To solidify the concepts presented throughout this tutorial, we use data from a randomized controlled trial conducted through the Adolescent Medicine Trials Network (study protocol 149; UCLA IRB #16-001674-AM-00006) as a motivating example. The study evaluated interventions to improve HIV prevention continuum outcomes in youth at high risk for acquiring HIV, as well as secondary outcomes including mental health symptoms, substance use, and housing insecurity. A detailed study description is found in Swendeman et al. (2019).

For this tutorial, we performed secondary data analyses using baseline data from the 1,486 adolescent study participants aged 14 to 24 years ($M = 20.89$, $SD = 2.15$). The adolescents varied in terms of gender identity, sexual orientation, race/ethnicity, and risk factors. As shown in Table 1, the outcome variable and many of the predictor variables had missing data. The percentage of missing values across the 47 variables varied between 0% and 12%. In total, 837 out of 47 × 1,486 = 69,842 (1.20%) observations were missing. Only 1,004 participants (68%) had complete observations on all 47 variables.

The research goal was to develop a predictive model for recent depression severity using a set of potential predictors (e.g., demographics, social determinants, risk factors, protective acts). The outcome variable was a continuous scale score from the nine-item Patient Health Questionnaire (PHQ-9; Kroenke et al., 2001) that indicated a participant's severity of depression symptoms in the past 2 weeks. The 46 candidate predictor variables included one nominal variable with four categories (i.e., race/ethnicity), 25 binary variables, and 20 continuous variables. To avoid convergence problems, some categories of variables were collapsed due to low cell counts (e.g., categories of gender identity were collapsed into only two categories: cisgender and transgender).

LASSO With Complete Data

LASSO Frameworand Loss Function—Ordinary least squares (OLS) estimation is a common way to obtain a solution for a regression model (i.e., assign values to the regression coefficients) with a continuous outcome. The optimal solution in OLS is determined by finding the values of the regression coefficients that minimize the following loss function

$$\sum_{i=1}^{N}\left(y_i - \left[\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right]\right)^2, \quad (1)$$

where N is the total number of participants, y_i is the raw score on the outcome for participant i, β_0 is the intercept, p is the total number of predictors β_j is the regression coefficient for predictor j, and x_{ij} is the raw score for participant i on predictor j. The terms in the brackets

can be replaced by $\hat{y}_i$, the predicted outcome score for participant *i* based on the model. Equation 1 is known as the residual sum of squares (RSS).

An advantage of OLS estimation is that it is the best linear unbiased estimator (BLUE). An estimator is unbiased if the sample estimates equal the population value in expectation. A disadvantage of OLS estimation is that the sample estimates have high variance such that from sample to sample, we can estimate vastly different parameter values. Models using OLS estimates typically have better model performance (e.g., predictive accuracy) in the sample used to estimate those coefficients compared with a different sample from the same population (McNeish, 2015). This is because OLS models are prone to overfitting. By overfitting we mean modeling relationships specific to the sample that do not exist in the population. Random noise in a sample can be confused as signal due to sampling error.

The LASSO is similar to OLS estimation, except it includes a penalized loss function that shrinks the coefficients toward zero and even sets some coefficients to exactly zero. For continuous outcomes, the formula the LASSO minimizes is

$$\sum_{i=1}^{N}\left(y_i-\left[\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}\right]\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|, \tag{2}$$

where the first half of the equation is the RSS and the last half of the equation is the shrinkage penalty. The RSS is small when the model has near perfect prediction of the outcome variable (i.e., the residuals are small). The shrinkage penalty, on the other hand, is small when the regression coefficients are near zero. The tuning parameter, λ (lambda), is a predetermined constant (we explain later how to determine it) that controls the impact of the shrinkage penalty on the parameter estimates. If λ is zero, then the solution is identical to OLS estimates. If λ is infinity, then all regression coefficients (except the intercept) are shrunk to zero. A tuning parameter in between these two extremes will fix some regression coefficients to zero and will estimate nonzero, but attenuated coefficients for the other coefficients. By doing so, the LASSO selects predictors that make sufficient contributions to predicting the outcome variable to achieve a parsimonious model compared to the OLS model.

The scaling of the predictors affects the magnitude of the coefficients, β_j, and thus impacts the shrinkage penalty and the solution. In other words, the scaling of predictors (e.g., measuring age in years vs. weeks) can lead to a predictor being favored over others simply because of scaling and not because of its strength in prediction. Thus, before a LASSO is estimated, all predictors are rescaled so that they have equal variances (James et al., 2013). Centering the predictors is not necessary because it only affects the value of the intercept, which is not penalized, but is often done to standardize each predictor to have a mean of zero and a standard deviation of one (e.g., Stata, StataCorp, 2019; and the glmnet package in R, Friedman et al., 2020, do this by default).

An advantage of the LASSO over OLS is that it decreases the risk of overfitting the model. By reducing the risk of overfitting, the LASSO has greater predictive ability compared to OLS (McNeish, 2015). However, reducing the risk of overfitting comes at the cost of adding

bias to the coefficients. By intentionally attenuating the coefficients, the LASSO sample estimates no longer equal the population value in expectation; however, the variances of the coefficient estimates from sample to sample are reduced. OLS estimates are unbiased; however, they have high variance from sample to sample. This is known as the bias-variance tradeoff.

Unlike OLS, there is no closed-form solution for the LASSO even for a fixed λ (Hastie et al., 2009). There are many options to obtain a solution for LASSO including coordinate descent (Daubechies et al., 2004; Wu & Lange, 2008), least-angle regression (LARS; Efron et al., 2004), and proximal gradient methods (Chen et al., 2012; Nesterov, 2005). For this tutorial, we used coordinate descent, the estimation procedure of the glmnet package in R (Friedman et al., 2020).

Before describing the steps of creating a LASSO model, we highlight a debate about how to best represent the categories of a categorical predictor variable with more than two categories (Huang & Montoya, 2020). One option is to overparameterize the model so that a reference group is not entered into the model. Specifically, if a variable has *c* categories, then *c* dummy codes are included in the LASSO (StataCorp., 2019, p. 190). This creates a singular (and non-invertible) design matrix. A singular design matrix cannot be used in OLS estimation because OLS requires inverting the design matrix. Thus, there is no OLS solution if the design matrix is singular. We can, however, estimate a LASSO solution with a singular design matrix by choosing which categories improve prediction and fixing the dummy codes for the remaining categories to have a coefficient of zero. Thus, we used *c* dummy variables to represent categorical variables with more than two categories for all LASSOs. Specifically, the four race/ethnicity categories were represented by four dummy variables. So even though there were 46 predictors initially (race/ethnicity treated as one variable), there were a total of 49 predictors used in the LASSO models (race/ethnicity entered as four variables) for this applied example.

Steps for Estimating a LASSO—In this section, we illustrate the process for estimating a LASSO assuming there is complete data (see Figure 1). Given missing values in the variables of interest in our applied example and the fact that a LASSO cannot be estimated with an incomplete data set, we used listwise deletion to remove participants with missing data on any of the studied variables (Figure 1, Step 1). This reduced the sample size in the applied example from 1,486 to 1,004. We used R software v. 3.1.2 (R Core Team, 2018) to analyze the data and the package glmnet (Friedman et al., 2020) to estimate the LASSO models. Annotated R code used for all approaches is available on the first author's OSF account.[1]

Split Data Into Training and Test Sets.: Machine learning methods like the LASSO tend to be exploratory, so an important step is to validate the final model in a holdout sample to assess the model's generalization error (Chen & Wojcik, 2016). A holdout sample can be synthetically created by splitting the observed data set into two nonoverlapping sets: a training set of data and a test set of data (Figure 1, Step 2). All models and modifications

[1] https://osf.io/7ys4m/.

to the models are analyzed in the training set. In the case of LASSO, this includes running multiple LASSOs with different values for the penalty parameter (lambda) to find the optimal value. Then, once the final model or models are decided on, they are evaluated in the test set. The test set should only be used at the very end of the analysis to provide a measure of prediction error of the model(s) on new data. Modifications to a model should not be made once fit to the test set. If a model performs well in the training set but does not perform well in the test set, then the model does not exhibit good predictive ability.

Selecting the ratio of the split of the training and test sets is complex due to competing needs: the need to have a large enough sample size in the training set to maximize performance of the statistical model and the need to have a large enough sample size in the test set to minimize the validation error (Guyon, 1997; James et al., 2013, p. 180). There are no clear guidelines on how to select an appropriate ratio given factors like the sample size, the ratio of signal-to-noise, and the complexity of the models being evaluated (Hastie et al., 2009). A common ratio is 75:25 such that 75% of the data is in the training set and 25% of the data is in the test set (Hastie et al., 2009, p. 222). If the sample size is relatively small, larger splits (e.g., 90:10) are recommended to accurately train the model (Dobbin & Simon, 2011; James et al., 2013). Because the sample size for a complete case analysis in the applied example is large ($N = 1{,}004$), we used the 75:25 split such that 753 participants were assigned to the training set and 251 participants were assigned to the test set.

***k*-Fold Cross-Validation in Training Set.:** To estimate the function in Equation 2, we need to obtain a λ value (Figure 1, Step 3). However, it is difficult to know which value of λ will produce the best prediction a priori. The goal is to choose a λ value that creates an interpretable model (i.e., shrinks some coefficients to zero), but not shrink coefficients so much that excessive bias is added into the estimates and the prediction error increases. There are a few ways to determine which λ value to use. We focus on *k*-fold cross-validation, a type of resampling method that improves the replicability of the model (James et al., 2013), because it optimizes the value of λ such that the predicted error in an independent sample is minimized. The eight steps of *k*-fold cross-validation are as follows:

CV1. Select a set of l candidate λ values.

CV2. Divide the data set into k roughly equally sized portions or folds.

CV3. Hold out the first fold as a validation sample.

CV4. With the remaining $k - 1$ folds, estimate a LASSO for every single candidate value of λ and save the coefficients for each of these l models.

CV5. Test each of these l models in the validation sample separately. Record a model performance measure like the mean squared error (*MSE*; i.e., the mean of Equation 1).

CV6. Repeat steps CV3, CV4, and CV5 so that each of the k folds acts as a validation sample one time. After this step is completed, there will be a total of $k \times l$ model performance measures.

CV7. To obtain one model performance measure for each λ, average the *k* measures of model performance for each candidate value of λ.

CV8. Select the value of λ associated with the best measure of model performance (e.g., smallest *MSE*).

The function cv.glmnet in R, which we use to select λ in this tutorial, conducts the above eight steps automatically (Friedman et al., 2020), but we explain each step in detail using the applied example for clarity.

The first step of *k*-fold cross-validation (i.e., step CV1 in the checklist) is to create a set of candidate values for λ. The default of cv.glmnet is to select 100 values of λ (i.e., $l = 100$) between λ_{min} and λ_{max}. The largest candidate value of λ, λ_{max}, is the smallest data-derived value that induces all coefficients to shrink to 0. In the case of our applied example, $\lambda_{max} = 4.465$. To confirm, we estimated a LASSO in the entire training set using λ = 4.465 and all regression coefficients except the intercept were shrunk to 0. The smallest candidate value of λ, λ_{min}, by default is .0001*λ_{max} if *N* is greater than *p*, which is the case for our example ($N = 753 > p = 49$). Based on these defaults, we would expect $\lambda_{min} = .0004$ for the applied example; however, due to a stopping rule in the cv. glmnet function, only 77 λ values between λ = .004 and λ = 4.465 were included in the set of candidate values. This stopping rule was put in place to reduce computation time. Otherwise, there are no drawbacks to testing 100 values between λ = .0004 and λ = 4.465.

The next step is to choose a value of *k* and divide the data set into *k* equally sized folds (i.e., Step CV2). There are many options for which value of *k* to choose. Leave-one-out cross-validation is when $k = N$. In this case, a single participant acts as the validation sample and the model is trained on the remaining observations. Other popular choices are $k = 5$ and $k = 10$ (Hastie et al., 2009). There is a bias-variance tradeoff when selecting the value of *k* (Yarkoni & Westfall, 2017). As *k* increases, the bias of the test error estimates decreases because the model is trained on more observations, but the variance of the test error estimates increases (James et al., 2013). Additionally, as *k* increases, the cross-validation procedure becomes more computationally demanding. For our analyses, we specified 10-fold cross-validation for all approaches. Dividing the training set of 753 cases into 10 equal folds resulted in seven folds with 75 participants and three folds with 76 participants. The participants need to be randomly assigned to these 10 folds. To reproduce the results, we set a seed so that participants would be assigned to the same fold each time the models were analyzed. If participants are shuffled into different folds, then a slightly different optimal λ value will be selected due to sampling error.

In Steps CV3–CV5 in the checklist, one fold is selected as a validation sample and the remaining folds are used to train the model. For instance, in our applied example, the first fold with 75 participants is labeled as the validation sample and the remaining nine folds are combined to form a sample of 678 participants. Then, 77 LASSOs are estimated using the 77 candidate λ values and these 678 participants. Next, the 77 estimated LASSOs are fit to the validation sample of 75 participants and the *MSE* is calculated for each model.

As it states in Step CV6, Steps CV3–CV5 are repeated so that each fold acts as a validation sample and the data set the LASSO is trained on is systematically resampled (i.e., a different combination of nine folds are merged into one data set to estimate the LASSOs). After this iterative process is completed, each λ value is associated with 10 different *MSE* values produced from the 10 different folds.

The 10 *MSE* values associated with a particular λ value vary slightly due to sampling variability (different coefficients due to different participants in the 9 folds and different participants in the validation fold). Thus, in Step CV7, the 10 *MSE* values are averaged for each λ value. For instance, in the applied example, the averaged *MSE* value for $\lambda = 4.465$ was 34.060 and for $\lambda = .004$ the averaged *MSE* value was 11.368.

The final step of cross-validation (i.e., Step CV8) is to select the λ value that produced the smallest averaged *MSE* value. The smallest averaged *MSE* value in our example was 10.667, which was produced by $\lambda = .143$. This λ value is referred to as the optimal λ value or the cross-validated λ value as it was selected via cross-validation.

Estimate a LASSO in Training Set.: Once the optimal λ value has been selected via *k*-fold cross-validation, a LASSO is estimated in the entire training set using the cross-validated λ value (Figure 1, Step 4). Returning to the applied example, we fit a LASSO with $\lambda = .143$ to the training set using the function glmnet (see annotated code for specific details). The estimated coefficients from this model are shown in the second column in Table 2. Using this model, a measure of model performance (e.g., *MSE*) can be calculated, though it is not necessary. The goal of the LASSO is not to determine how well the model predicts outcomes in the training sample, but in the test sample.

Fit Estimated LASSO Model to Test Set.: The training model (i.e., the coefficients in Table 2) is then fit to the test set to produce predicted outcome scores for participants in the test set (Figure 1, Step 5). Using the predicted and observed outcome scores, a model performance measure (e.g., *MSE*) is calculated (Figure 1, Step 6). This estimate of the test error rate quantifies the generalizability of the model to future samples. If multiple models are compared (e.g., LASSO vs. OLS), the model associated with the lower test *MSE* is typically selected as the better model. If only one model is examined, comparing the test *MSE* to the training *MSE* gives a sense of the generalizability of the model; however, there are no guidelines for what constitutes a concerning discrepancy. As shown in Table 2, the test *MSE* for the applied example was 12.579. This value is larger than the training *MSE* (10.085), which is often the case.

Perspectives on Statistical Inference When Using LASSO—There is no consensus on how to derive standard errors for LASSO coefficients (Kyung et al., 2010). Part of the complication is that implementations of LASSO often return a value of zero for the associated standard error if the LASSO coefficient is set to zero. Proposals for calculating standard errors include bootstrapping (Chatterjee & Lahiri, 2011; Tibshirani, 1996; Wan et al., 2015), using the standard errors of the coefficients estimated via ridge regression as an approximation of the standard errors of the coefficients estimated via LASSO (Tibshirani,

1996), using a sandwich estimator (Fan & Li, 2001), and post-LASSO estimation (Belloni & Chernozhukov, 2013; Efron et al,. 2004; Hansen, 2016).

Given the complication with standard errors and because calculating the degrees of freedom is not straightforward (Zou et al., 2007), any attempt to calculate *p*-values is necessarily complex. Meinshausen and Bühlmann (2010) and Wasserman and Roeder (2009) proposed methods for calculating *p*-values that subset the data set, which can create complications for small sample sizes. Lockhart et al. (2014) created significance tests that utilize the full data set using results from the LARS algorithm. A few LASSO-based algorithms, which are available in Stata software (StataCorp., 2019), have been proposed to estimate regression coefficients, standard errors, and *p*-values for a specified subset of predictors and to select from potential control variables that are entered in the model (Belloni et al., 2012; Belloni et al., 2014a, 2014b; Belloni et al., 2016; Chernozhukov et al., 2018; Chernozhukov et al., 2015).

We view the relevance of LASSO coefficient inference as dictated by the purpose of the analysis. More broadly, it has been noted that statistical significance of the coefficients of candidate predictor variables does not necessarily imply good predictive ability in future samples (Lo et al., 2015). Alternatively, nonsignificant independent variables may be important for prediction. Here, we view the central goal of the LASSO as creating a predictive model in a way that avoids the risk of overfitting and in contexts where inferential or causal claims are not of primary importance (Yarkoni & Westfall, 2017). Thus, we focus on the predictive ability of fitted models rather than the significance of the coefficients of the selected variables.

Multiple Imputation

General Framework—Rather than removing participants with missing data, MI can be used to handle the missing values as it can improve the predictive ability of a machine learning model compared with listwise deletion (Poulos & Valle, 2018). MI consists of three phases: imputation, analysis, and pooling (Enders, 2010; Little & Rubin,2019; Rubin, 1987; Schafer, 1997; van Buuren, 2018). In the imputation phase, multiple copies of the data are created where the missing values are replaced with plausible values drawn independently from an appropriate statistical model. In the analysis phase, the resulting imputed data sets are analyzed separately using statistical methods applicable to complete data. Finally, in the pooling phase, the parameter estimates and standard errors obtained from each imputed data set are combined using Rubin's rules (Rubin, 1987) for a single set of results that support an overall inference.

The overall MI estimate of a parameter is the average of the parameter estimates over the multiply imputed data sets, and the variance of the MI parameter estimate incorporates both within-imputation variability (the sampling variation of the estimate in each imputed data set) and between-imputation variability (the variation in estimates between the data sets) of the estimates. Thus, MI takes into account the uncertainty in the estimate due to the missing data (Little & Rubin, 2019). Standard implementations of MI are valid under the unverifiable assumption of missing at random (MAR), where the probability of a value being missing depends on the other observed data but not on the unobserved data (Little & Rubin, 2019).

Joint Modeling and Fully Conditional Specification Estimation Strategies— Two general estimation frameworks for implementing MI include joint modeling (JM; Schafer, 1997, 1999; Schafer & Olsen, 1998) and fully conditional specification (FCS; van Buuren, 2018; van Buuren et al., 1999; van Buuren et al., 2006), also known as multivariate imputation by chained equations (MICE; Carpenter & Kenward, 2013) or sequential regression (Raghunathan et al., 2001). Both the JM and FCS approaches are well represented in widely available statistical software. JM draws missing values for all incomplete variables from an explicit multivariate distribution, often through the use of a Markov chain Monte Carlo (MCMC) algorithm that includes steps involving overlapping conditional distributions (Jackman, 2000; Tanner & Wong, 1987). Within a multivariate normal model, for example, the model parameters are represented through a vector of means and a variance-covariance matrix, and the conditional distributions are linear regressions. The iterative sampling procedure begins with (a) estimating the model parameters for all the variables included in the imputation model conditional on the observed data, (b) drawing missing values from a predictive distribution that conditions on the current parameter estimates, and (c) drawing new values of the parameters from their posterior distribution given complete data. Steps (b) and (c) are iterated until a convergence criterion is satisfied.

FCS similarly draws missing values iteratively from a specified set of overlapping (usually univariate) conditional distributions for each incomplete variable, conditioning on the remaining variables. In particular, FCS sets up a sequence of regression models where each incomplete variable is regressed on all others, and the iterative algorithm proceeds by estimating parameters of each regression model one at a time. FCS procedures are understood not as rigorous implementations of MCMC procedures but rather as approximations motivated by MCMC procedures that tend to have satisfactory statistical properties.

In general, the JM imputation approach has a secure theoretical foundation through assuming a parametric model for multivariate data and can handle mixtures of continuous and categorical variables. This framework allows us to accommodate incomplete binary, ordinal, and nominal variables via underlying normal latent variables (Muthén & Muthén, 1998–2017; Quartagno & Carpenter,2019; Quartagno & Carpenter, 2020). Unlike JM imputation, FCS is very flexible in allowing an appropriate univariate regression specification for each incomplete variable, and in accommodating mixed types of missing data. However, there is the theoretical issue of incompatibility between specified conditional distributions for incomplete variables. In many settings, the impact of incompatibility among conditional distributions is apt to be relatively minor (Raghunathan et al., 2001; van Buuren, 2018; van Buuren et al., 2006); greater concerns are apt to be present in scenarios where the analysis model includes interaction terms or nonlinear effects involving incomplete variables. Model-based (fully Bayesian) imputation methods (Enders et al., 2020; Erler et al., 2016; Ibrahim et al., 2002; Kim et al., 2018; Kim et al., 2015; Ludtke et al., 2020) and an extension of the FCS approach known as substantive model-compatible imputation (Bartlett et al., 2015) are alternative strategies that have been recently developed to tailor imputations around a specific model that accommodates incomplete interactive or nonlinear effects.

In the case of imputation LASSO approaches, the same set of candidate predictors (in the same form) used in the LASSO should appear in the imputation model. Therefore, the final LASSO model will be nested within the imputation model. Excluding any candidate predictor from the imputation model may introduce additional bias to the parameter estimates separate from the added bias due to attenuation (e.g., set a coefficient incorrectly to zero) because the imputation model does not preserve the associations between the imputed values and excluded variable. Furthermore, any potential auxiliary variables (i.e., variables that are not included in the analysis model but predict variables with missing data or predict the mechanism giving rise to missing data) need to be included in the imputation model (Collins et al., 2001; Graham, 2012). Strong auxiliary variables can enhance the plausibility that the missing data mechanism is MAR, help improve the imputation process, and increase power.

MI Details for Applied Example—We used the Blimp 2.1 application (Keller & Enders, 2019) and adopted MI using FCS to handle missing values in our applied example. Annotated Blimp code is available on the first author's OSF account.[2] For simplicity, our analysis model did not have interaction effects nor nonlinear terms and no auxiliary variables were included in the imputation model. Before imputing the missing values, the number of imputations (m) needs to be selected. Recent literature recommends that the number of imputations should at least be greater than the percentage of missing data in the analysis variables (White et al., 2011) and some recommend 100 imputations or more to replicate standard errors and increase power (von Hippel, 2018). For our applied example, we generated $m = 50$ imputed data sets to achieve a precise inclusion frequency (explained later) and better precision in point estimates (von Hippel, 2018).

Convergence of the MCMC algorithm was determined by calculating potential scale reduction factors (PSRF; Brooks & Gelman, 1998; Gelman et al., 2014) for each parameter and examining trace plots of the parameters, which are the plots of estimated parameter values against the MCMC iteration numbers, to evaluate mixing of the Markov chains. A rule of thumb is to conclude the algorithm converged if all PSRF values are below 1.10 (Gelman et al., 2014). Using the applied example, the largest PSRF value after 800 iterations was 1.08. Thus, we used a conservative burn-in period of 1,000 iterations and a thinning interval of 1,000 iterations to generate 50 imputed data sets.

Imputation LASSO Approaches

Figures 2, 3, and 4 illustrate the steps for implementing the separate approach, stacked approach, and MI-LASSO, respectively. The first two steps in Figures 2–4 (i.e., using multiple imputation to generate m imputed data sets and splitting each imputed data set into a training and test set) are the same across the three imputation LASSO approaches. For multiply imputed data, each row of data for the same participant should be in the same split. Thus, for the running example, we used a 75:25 ratio for each imputed data set to split the 1,486 participants such that the same 1,114 participants were always assigned to the training set and the remaining 372 participants were always assigned to the test set. For

[2] https://osf.io/7ys4m/.

simplicity, we refer to these data sets collectively as the imputed training sets and imputed test sets, respectively. The total number of imputations (m) is preserved as there are 50 imputed training sets and 50 imputed test sets for the applied example. We now describe the remaining steps that slightly differ across the three imputation LASSO approaches.

Separate Approach

The separate approach (see Figure 2) involves fitting a variable selection procedure separately in each of the m imputed data sets. This will typically result in different variables selected in each imputed data set. There is debate as to how to determine the final selection of variables and how to pool the coefficients if different variables are selected (Zhao & Long, 2017). The first discussion of the variable selection problem within the MI framework was raised by Brand (1999), who proposed a two-step solution using stepwise regression. Others expanded on Brand's work by exploring alternative solutions to the variable selection problem using stepwise regression (Wood et al., 2008), LARS (Lachenbruch, 2011), LASSO (Thao & Geskus, 2019), and Bayesian methods (Yang et al., 2005). When using the LASSO as the variable selection procedure for the separate approach, the steps described in the Steps for Estimating a LASSO section can be used as is except they are applied to each imputed data set individually.

k-Fold Cross-Validation in Imputed Training Sets—The eight steps of k-fold cross-validation (i.e., CV1–CV8) are conducted in each imputed training set separately (Figure 2, Step 3). This will result in m different λ values. To our knowledge, there has not been an investigation as to the best way to select the cross-validated λ value for multiply imputed data. The only explicit reference to validating λ using the separate approach was made by Thao and Geskus (2019). They performed 10-fold cross-validation on just one imputed data set to obtain a single cross-validated λ value. For this tutorial, we explored five other options: (a) allow each imputed training set to have its own imputation-specific λ value, or use the (b) mean, (c) median, (d) minimum, or (e) maximum of the m λ values for all imputed training sets. For the applied example, the cross-validated λ values for options (b–e) were .109, .112, .078, and .136, respectively. These five options led to a different selection of variables when including any variable selected in at least one imputed training set. Option (d), the least penalized option, selected 33 predictors; option (a) selected 30 predictors; options (b) and (c) selected the same 29 predictors; and option (e), the most penalized option, selected 24 predictors. For simplicity, we selected option (b), λ = .109, as the cross-validated λ value used for all imputed data sets as a compromise between the extreme values.

Estimate LASSO Models in Imputed Training Sets—Using the cross-validated λ value(s) from the previous step, a LASSO is estimated in each imputed training set (Figure 2, Step 4), resulting in m LASSO results (henceforth referred to as imputation-specific LASSO models). The set of selected predictors varies across the imputation-specific LASSO models: some predictors are never selected, some predictors are selected in some of the imputed training sets, and some predictors are selected in every imputed training set. This complication has resulted in debate on the appropriateness of pooling and how to pool coefficients and model performance measures if it is appropriate.

Before discussing the mathematics of pooling, it is important to address if pooling is appropriate when different variables are selected across the imputed data sets. Chen and Wang (2013) argue that if different predictors are selected, then the coefficient estimates cannot be pooled because "covariates of regression models in each imputation are different" (p. 3649). For example, if the variable ever smoked was selected in some but not all the imputed data sets, then the partial regression coefficients for the other selected variables will account for the effect of ever smoked in some of the imputed data sets but not in others. However, if the LASSO was estimated without including the nonselected predictors as inputs in the model, the resulting coefficient estimates will differ from the original estimates. Thus, the nonselected predictors are in effect controlled for. Regardless, this appears to be a negligible issue because the goal of the LASSO is to create a predictive model, not inference. Therefore, for the applied example, we proceed assuming that pooling the regression coefficients is acceptable even when different covariates are selected across the imputed data sets.

Previous studies obtained the pooled estimate of a slope for a predictor by averaging the m LASSO coefficients in the m imputation-specific LASSO models for the predictor (Musoro et al., 2014; Thao & Geskus, 2019). This is what is done for OLS estimates of regression coefficients as it is assumed that the estimates form a normally distributed sampling distribution. In the case of LASSO estimates, however, this may not be the accurate pooling procedure. There are three scenarios for the combination of individual slope values: (a) the pooled slopes reflect nonzero slopes in all m imputed data sets, (b) the pooled slopes reflect a mix of zero and nonzero values, and (c) the pooled slopes reflect zero slopes in all m imputed data sets. Returning to our example, Table 3 shows the 50 slope estimates for four predictors in the applied example. Anxiety as measured by the GAD-7 was selected in all 50 imputed data sets, ability to make friends was selected in 37 imputed data sets, ever smoked was selected in 23 imputed data sets, and age was not selected in any imputed data sets. Both ability to make friends and ever smoked variables reflect scenario (b) such that the pooled estimates reflect a mixture of zero and nonzero values. In this scenario, because there is an inflation of zero slope values and all nonzero slope values have the same sign (e.g., all nonzero slope values are positive for the ever smoked variable and all nonzero slope values are negative for the ability to make friends variable), it is difficult to believe that the sampling distribution of the slopes is a normal distribution. In the absence of guidance on how to pool results given this inflation, we averaged the 50 LASSO coefficients for all predictors. The coefficients for the final model of the applied example are presented in Table 2 under the Separate approach column. The final model is the predictive LASSO model that uses the MI point estimates for the coefficients of selected variables. In this case, the set of variables selected in the final model are any predictors that were selected in at least one imputed training set.

A model performance measure such as the *MSE* can be calculated in the training set to compare it to a model performance measure in the test set. In the context of multiple imputation, this means calculating the *MSE* in each imputed training set using the corresponding imputation-specific LASSO model and pooling the m *MSE* values by taking the average. Using the applied example, the 50 *MSE* values ranged from 10.42 to 11.00 with an average value of 10.68.

Potential downsides of the process just described are the selection of noise variables (a predictor only needs to be selected in one imputed data set to be included in the final selection of predictors) and using different models (due to inconsistently selected predictors) to calculate a model performance measure. Thus, previous researchers have used an inclusion frequency or IF (i.e., number of imputed data sets the variable was selected in divided by the total number of imputed data sets) to select fewer variables. If the IF for a predictor is at or above a threshold value, π, where $0 < \pi \leq 1$ (Heymans et al., 2007), then the predictor is included in the final model. Common choices for π are (a) $\pi = 1/m$ (i.e., predictor was selected in at least one imputed data set); (b) $\pi = .5$ (predictor was selected in at least half of the imputed data sets); and (c) $\pi = 1$ (predictor was selected in all of the imputed data sets; Thao & Geskus, 2019; Vergouwe et al., 2010; Wood et al., 2008). Not surprisingly, the final selection of variables is sensitive to the value of π. Strategy (c) leads to a more parsimonious model relative to any other choice for π as fewer variables are included in the final selection of predictors. Strategy (a), which is the process we just described, is more susceptible to selecting noise variables, especially as the number of imputations (m) increases (Thao & Geskus, 2019), but the estimates have less bias compared to larger π values. Thus, the selection of π is related to the bias-variance tradeoff such that as π increases, the bias of the estimates increases (more shrinkage occurs by virtue of not selecting predictors), but the variance of the estimates decreases.

Returning to the applied example, estimating LASSOs in the imputed training sets using the mean cross-validated λ value (.109) led to 29 variables selected in at least one imputed training set, 20 variables selected in at least 50% of the imputed training sets, and 15 variables selected in all imputed training sets. The inclusion frequencies for all predictors are provided in Table 2.

If using $\pi = .5$, which is a popular choice for applied and simulation studies (Lachenbruch, 2011) as it balances the bias-variance tradeoff, the coefficients for predictors selected in fewer than 50% of the imputation-specific LASSO models need to be manually changed to 0 when calculating model performance measures like the *MSE* Returning to Table 3, ever smoked would not be included in the final selection of variables in this scenario because it was selected in only 46% of the data sets. Thus, its coefficients in the imputation-specific LASSO models would be changed to 0 and the m MSEs would be calculated using these m revised imputation-specific LASSO models. Additionally, the coefficient in the final model associated with ever smoked would be 0. This is illustrated in the second to last row of Table 3. If $\pi = 1$, the last row of Table 3 shows that of the four variables, only anxiety would have a nonzero coefficient in the final model (and in each revised imputation-specific LASSO model) because it was the only variable consistently selected.

Fit Estimated LASSO Models to Imputed Test Sets and Pool Results—There is little to no guidance on how to fit the training models in the imputed test sets (Figure 2, Step 5). Should the final model (using the pooled estimates) or the imputation-specific LASSO models be fit to the test sets? In an applied example, Musoro et al. (2014) calculated model performance measures in each imputed test set using the coefficients from the final model, but it is unclear if this method gives the best estimate of the test error rate. To be consistent with how the pooled *MSE* was calculated in the training set, we fit each imputation-specific

LASSO model to its corresponding imputed test set and calculated the *MSE* (Figure 2, Step 6). We then calculated the pooled test *MSE* (Figure 2, Step 7) by taking the average of the 50 *MSE* values. The 50 *MSE* values ranged from 11.11 to 11.76 with an average value of 11.37. This value can be compared with the training *MSE* (10.68) to get a sense of the generalizability of the model, but there are no guidelines for determining what is an acceptable difference between the two values.

Stacked Approach

Rather than estimating a LASSO in each imputed data set, a LASSO model can be estimated using the stacked set of imputed data sets. A stacked data set is a long format data set that includes all the imputed data sets stacked on top of one another with $N \times m$ rows for N participants and m imputed data sets. In the stacked approach, the m imputed training sets are stacked on top of one another as are the m imputed test sets, resulting in a stacked training set (Figure 3, Step 2a) and a stacked test set (Figure 3, Step 2b), respectively. For the applied example, this results in a stacked training set with 1,114 × 50 = 55,700 rows and a stacked test set with 372 × 50 = 18,600 rows. Fitting the LASSO using the stacked approach is straightforward as there is one data set with no missing values. Because this approach requires only one analysis, there is not an inconsistent selection of predictors, simplifying the process and removing decisions needed in the separate approach such as selecting $\boldsymbol{\pi}$.

k-Fold Cross-Validation in Stacked Training Set—Steps CV1–CV8 of *k*-fold cross-validation can be conducted using the stacked training set (Figure 3, Step 3) as is with one potential exception: the inclusion of weights when estimating the LASSO. Generally, when an analysis is performed using the stacked data set, the parameter estimates are unbiased and consistent if they are unbiased and consistent for the individual imputed data sets, but the standard errors are underestimated due to the inflated sample size (Cohen et al., 2003; Zhao & Long, 2017). (For the applied example, the number of participants in the training set is 1,114, but the statistical program assumes a sample size of 55,700 when conducting analyses using the stacked training set.) Similar to the separate approach, the stacked approach was first applied using stepwise regression (Wood et al., 2008). Because standard errors are needed for that selection process, the standard errors were corrected by incorporating weights. Multiple weighting strategies have been proposed for stepwise regression including weights for individuals (Wood et al., 2008) and weights for variables (Vergouwe et al., 2010).

Weights have also been included when using machine learning variable selection methods. Wan et al. (2015) fit the elastic net on the stacked imputed data set utilizing one of two weighting schemes: $w_i = 1/m$ and $w_i = f_i/m$, where f_i is the number of predictor variables with no missing values for participant i (the authors assumed complete data on the outcome variable) divided by the total number of predictors. If a participant has complete data on all predictors, then for both weighting schemes, the weight for their row in one imputed data set is equal to $1/m$. By summing across the imputed data sets, the total weight for one participant with complete data is equal to 1. In their simulation study, both weighting schemes exhibited similar predictive performance. Thao and Geskus (2019) used

the LASSO to compare the weighting scheme $w_i = f_i/m$ to a stacked approach that did not use weights and found that both approaches had similar predictive performance. If standard errors are not calculated and only point estimates are desired, it is unclear why a weighting scheme is needed when fitting a LASSO as the stacked data set yields consistent estimates (van Buuren, 2018).

Returning to the applied example, we conducted 10-fold cross-validation in the stacked training set with no weights, with $w_i = 1/m$, and with $w = f_i/m$. The glmnet function in R easily incorporates weights in the analysis (see annotated code for details). The first two weighting schemes produced identical cross-validated λ values while the third weighting scheme produced a slightly different cross-validated λ value (different at the sixth decimal place). Note, the default option of the glmnet package in R is to standardize the predictors in relation to the stacked training set. This standardization retains between-cluster variation.

Estimate LASSO in Stacked Training Set—Using the chosen weighting scheme from the previous step, a LASSO is estimated in the stacked training set using the cross-validated λ (Figure 3, Step 4). In our case, the unweighted and $1/m$ weights produced identical point estimates and thus model performance measures (*MSE* = 10.4560). Using f_i/m weights led to one fewer predictor selected (self-rating of ability to live drug free) and slightly different coefficient estimates, but similar model performance measures (*MSE* = 10.4562). Because it has not been shown that weights are needed for the point estimates, we present the results for the unweighted solution only in Table 2. Only two predictors were not selected by the LASSO: the Latinx and ever homeless dummy variables. Because all other race/ethnicity dummy variables were selected, the Latinx group became the reference group in this model.

Fit Estimated LASSO to Stacked Test Set—Next, the training model is fit to the stacked test set (Figure 3, Step 5) and a model performance measure is calculated (Figure 3, Step 6). Although the f_i/m weights led to slightly better model performance measures in the test set (*MSE* = 11.642) compared with the other two weighting schemes (*MSE* = 11.647), we cannot recommend these weights over the unweighted solution in all situations. In fact, when using a different split for the training and test sets, the unweighted LASSO performed better in the test set.

The test *MSE* for the stacked approach should be compared to the test *MSE* for the separate approach. Because the test *MSE* for the separate approach is smaller than the test *MSE* for the stacked approach, we can conclude the separate approach is better than the stacked approach for prediction for this applied example.

MI-LASSO

To ensure consistent variable selection across the imputed data sets, a group LASSO can be applied to the stacked set of imputed data sets (Chen & Wang, 2013). The group LASSO enters a set of variables as a group (Yuan & Lin, 2006). This is sometimes seen when including an interaction and its main effects in the LASSO so that the main effects are selected if the interaction is selected or when a set of dummy variables that represent the same nominal variable (e.g., race) is entered into the LASSO. The group LASSO will select

all the variables within the group or shrink the coefficients for all variables within the group to zero.

The MI-LASSO approach uses the group LASSO penalty to estimate a LASSO in all imputed data sets jointly. Specifically, the MI-LASSO minimizes

$$\sum_{d=1}^{m}\sum_{i=1}^{N}\left(y_{di}-\left[\beta_{d0}+\sum_{j=1}^{p}\beta_{dj}x_{dij}\right]\right)^2+\lambda\sum_{j=1}^{p}\sqrt{\sum_{d=1}^{m}\beta_{dj}^2}, \quad (3)$$

which is similar to Equation 2; however, the coefficients are now indexed by imputed data set as well as by predictor. Like the separate approach, each imputed data set has its own intercept and set of p estimated regression coefficients. Unlike the separate approach, the m imputed data sets are analyzed jointly in one analysis rather than separately. Each predictor acts as its own group such that the m regression coefficients associated with one predictor will either all be zero or all be nonzero. This approach assumes that if a predictor is important, then it should be selected in all imputed data sets. Conversely, if a predictor is not important, then it should not be selected in any imputed data set. While intuitively this seems like an optimal feature of the approach, rendering a binary judgment on the importance of the predictor (i.e., selected or not selected) does not allow for uncertainty due to missing data. Figure 4 illustrates the process of implementing the MI-LASSO.

k-Fold Cross-Validation Via MI-LASSO—The authors of the MI-LASSO provide SAS code and an R function for using the MI-LASSO,[3] but rather than using cross-validation or a training/test split, the optimal λ value is the value that minimizes the BIC. Unlike the *MSE*, which decreases as model complexity increases (e.g., the number of predictors increases), the BIC balances model performance with model complexity. However, as stated earlier, calculating the degrees of freedom for the LASSO is not straightforward (Zou et al., 2007), calling into question the accuracy of the BIC, which relies on degrees of freedom. Thus, we modified the MI-LASSO function to calculate the *MSE* and conducted 10-fold cross-validation on the stacked training set (Figure 4, Step 3). As seen in Table 2, the cross-validated λ value for the MI-LASSO approach (λ = 27.542) was much larger compared to the other approaches. The reason for this is because the MI-LASSO is penalizing a different quantity compared to the other two imputation LASSO approaches. As shown in Equation 2 and 3, the number of coefficients minimized for the MI-LASSO is m times greater than the number of coefficients minimized for the LASSO.

Estimate MI-LASSO in Imputed Training Sets Jointly—The MI-LASSO is estimated using the cross-validated λ value and the imputed training sets (Figure 4, Step 4). The result is m sets of coefficients that vary across the imputed data sets. Unlike the separate approach, if the coefficient for a particular variable is zero in one imputed data set, then it is zero in all other imputed data sets due to the group LASSO penalty, resulting in consistent variable selection and no need for the arbitrary threshold parameter, π. However, the nonzero coefficients are not identical across imputed data sets. Thus, the coefficient

[3] www.columbia.edu/~qc2138.

estimates for a selected predictor need to be pooled to obtain a single point estimate. Like in the separate approach, the coefficients for a selected variable are pooled by calculating the mean (Musoro et al., 2014; Thao & Geskus, 2019). Estimating the MI-LASSO using the cross-validated λ value (27.542) led to the selection of 20 predictors, a subset of the 29 predictors selected by the separate approach. Table 2 presents the pooled point estimates for the regression coefficients in the final LASSO model.

A model performance measure like the MSE can be calculated in each imputed training set using the imputation-specific LASSO coefficients. To obtain a single point estimate, the m *MSE* values are pooled by taking the average. For the applied example, the 50 *MSE* values ranged from 10.45 to 11.01, with an average value of 10.70.

Fit Estimated LASSO Models to Imputed Test Sets and Pool Results—The final steps of the MI-LASSO procedure involve fitting the model in the imputed test sets (Figure 4, Step 5) and calculating a pooled model performance measure (Figure 4, Steps 6, 7); however, like the separate approach, it is not clear if the final model or the imputation-specific LASSO models should be tested. To be consistent with the separate approach, we calculated the *MSE* values using the imputation-specific LASSO models. The 50 *MSE* values ranged from 11.14 to 11.74, with an average value of 11.37. Comparing this value to the test MSEs of the separate and stacked approaches, the separate approach is likely to perform the best in a holdout sample and thus should be selected as the best predictive model of the three imputation LASSO approaches.

Discussion

This article provides a tutorial on how to implement three approaches for estimating a LASSO with multiply imputed data. In doing so, we moved the needle on an analytic framework that bridges the gap between the machine learning and traditional statistical "cultures" as they were referred to in Leo Breiman's landmark paper (Breiman, 2001). Breiman's paper, as well as numerous other articles (e.g., Donoho, 2017; Mukhopadhyay & Wang, 2020; Yarkoni & Westfall, 2017), have advocated for a blended culture so that the best method for a given task is chosen, such as LASSO for increasing the predictive ability of a model in future samples. The presence of missing data, a common and practical issue in psychological analyses, often supersedes conceptual reasons for selecting analytic approaches. Traditional statistical analyses such as standard linear regression are favored because the theoretical basis for combining them with multiple imputation are more straightforward and more easily implemented through available statistical software relative to machine learning methods.

A strength of this tutorial was the provision of step-by-step procedures that can be applied to free software packages that are accessible by psychologists and other researchers. Until now, there has not been a clear discussion on how to validate the LASSO when using multiply imputed data. We provided clear guidance on how to conduct cross-validation and include a training/test split for the approaches discussed. One decision point for determining which imputation LASSO approach to implement are the available options across different statistical software packages. In this vein, the separate and stacked approaches are ideal

because they can be implemented in freely-distributed software packages (e.g., Blimp and glmnet were used for MI and LASSO in this article, respectively), as well as commercial software that supports MI and LASSO (e.g., Stata). Data preparation and analysis is easiest to conduct using the stacked approach relative to the other two approaches as LASSO can be applied to a single stacked data set. To the best of our knowledge, the MI-LASSO approach can only be implemented via the R or SAS code provided by the original authors, making options limited (e.g., *k*-fold cross-validation automation unavailable, BIC is the only model performance measure provided, elastic net would need to be programmed, code needs modification if using a categorical outcome). Additionally, it is the most computationally intensive of the three procedures. In our analyses, MI-LASSO took over 10 min to conduct 10-fold cross-validation for two λ values whereas it took 6 s for the separate approach to conduct 10-fold cross-validation for 77 λ values.

Another deciding factor in selecting an imputation LASSO approach is model interpretability. In the separate approach, each imputed data set can have a different set of selected variables, bringing into question how valid it is to pool parameter estimates if they have different interpretations across the different imputed data sets (the partial regression coefficients are controlling for the effects of different predictor variables). The stacked approach avoids this issue but, in our example, all but two of the 49 predictor variables were selected, which makes for a less parsimonious model compared with the other approaches. Additionally, the stacked approach had the worst predictive ability (i.e., largest test set *MSE*). The inflated sample size of the stacked data set may affect the value of λ, leading to an inaccurate number of selected variables. With almost all predictors selected due to a small cross-validated λ value, it is not surprising to see that the test *MSE* was the worst compared to the other two imputation LASSO approaches. In our example, the MI-LASSO is the most intuitively appealing of the three imputation LASSO approaches because it does not require selecting an arbitrary threshold, only one analysis was required after cross-validation, and the final model selected fewer variables compared with the stacked approach.

It is difficult draw a clear winner from the three imputation LASSO approaches, but the stacked approach performed the worst of the three based on the test model performance measures. However, it is important to provide caution in generalizing these findings as the analyses were applied to a single data set. Prior studies did not find any difference in the number of variables selected between the stacked and separate approaches (Thao & Geskus, 2019; Wood et al., 2008). This could be because our results are an outlier (e.g., we used $m = 50$ while others tend to use smaller values like $m = 5$ or $m = 20$) or because the data structure of our applied example was not examined in their simulation studies. If the latter, simulation studies need to expand the examined simulation models/conditions to increase the generalizability of the findings. Simulation results from prior studies have yet to confirm a preferred imputation LASSO approach, especially in striking a balance between improvements in performance across multiple metrics and low implementation burden.

In this tutorial, we illustrated the application of the separate, stacked, and MI-LASSO approaches for fitting a LASSO to multiply imputed data due to their ease of implementation through readily available software and/or theoretical basis. Other approaches to explore include a multiple imputation random LASSO (MIRL) method (Liu et al., 2016) that

combines MI and random LASSO (Wang et al., 2011), methods that perform variable selection on data sets that combine MI and bootstrapping (e.g., Deng et al., 2016; Long & Johnson, 2015; Musoro et al., 2014; Thao & Geskus, 2019; Zhao & Long, 2016), group LASSO methods that use penalized pooled objective functions that force consistent variable selection across the imputed data sets (e.g., Du et al., 2020; Geronimi & Saporta, 2017; Marino et al., 2017), generating only one imputed data set, and Bayesian penalized regression techniques (Makalic & Schmidt, 2016).

We stated earlier that the imputation LASSO approaches discussed in the tutorial could easily be applied to elastic net because they both select predictors by fixing coefficients to zero. Ridge regression, on the other hand, is a regularized regression method that does not fix coefficients to zero and thus does not suffer from the same complications discussed in the tutorial. Further, ridge regression has a closed-form solution, making the use of inferential statistics more tenable. While our tutorial was specific to regularization of the standard linear regression model, the general framework of the three imputation LASSO approaches can be applied to more complex statistical models that incorporate regularization such as regularized structural equation modeling (Liang & Jacobucci, 2020) and regularized partial correlation networks (Epskamp & Fried, 2018). Additionally, future research should explore how these frameworks can be applied to other machine learning and artificial intelligence algorithms—such as random forests, decision trees, and artificial neural networks—encountered in psychosocial studies (Feelders, 1999; Parker, 2010; Poulos & Valle, 2018; Rodgers et al., 2021; Twala et al., 2008).

There are further extensions to explore that our tutorial did not cover. For instance, our analyses explored linear relationships between predictors and a continuous outcome using cross-sectional data. Other psychological studies call for evaluations of nonlinear relationships (e.g., interaction effects) between predictors and discrete outcomes using longitudinal data. Most existing approaches for variable selection in the presence of missing data are developed under the MAR mechanism, which is often implausible in many settings. Although it is not widely done in practice, MI can accommodate known missing not at random (MNAR) mechanisms (i.e., the probability of a value being missing depends on the unobserved values of that variable) under the selection modeling (Beesley & Taylor, 2021; Carpenter et al., 2007; Hayati Rezvan et al., 2015) and pattern-mixture modeling (Hayati Rezvan et al., 2018; Leacy et al., 2017; Tompsett et al., 2018; Tompsett et al., 2020) frameworks. Addressing challenges that arise during implementation of variable selection strategies when using MI to address MNAR missingness is an area for future development. Further work could explore the validity of inferential statistics for LASSO with multiply imputed data. Despite the various methods for calculating standard errors for the LASSO, significance tests have not yet been applied for the LASSO or elastic net in the context of MI.

This tutorial showcased and evaluated three imputation LASSO approaches to prompt a wider scale adoption of machine learning variable selection approaches by psychological researchers, even when the variables have missing values. Given the pros and cons of each imputation LASSO approach, we refrain from recommending one over another and instead provide steps for researchers to use each imputation LASSO approach. The intersection of

the missing data literature and machine learning literature is relatively small (Twala et al., 2008), but this article contributes to that intersection, illustrating places for further research and providing guidance for methodological and applied researchers on how to use three different approaches for fitting a LASSO when using multiple imputation to handle missing data.

Acknowledgments

Grants from the National Institute of Mental Health supported the Heather J. Gunn and Panteha Hayati Rezvan (T32MH109205) and W. Scott Comulada (P30MH058107). Heather J. Gunn was also supported by a grant from the Mayo Comprehensive Cancer Center Grant (P30CA15083-47). The data come from an National Institute of Child Health and Human Development-funded U19 Cooperative Agreement for the Adolescent Medicine Trials Network (ATN) (U19HD089886).

The data set used as an example for this tutorial was obtained from study protocol 149 of the ATN Comprehensive Adolescent Research and Engagement Studies (ATN CARES; University of California, Los Angeles IRB 16-001674-AM-00006; Swendeman et al., 2019). The data set and syntax used for the analyses can be found on Heather J. Gunn's OSF account (https://osf.io/7ys4m/). The analyses described in this tutorial were presented at the annual International Meeting of the Psychometric Society on July 21, 2021, at Mayo Clinic's Statistical Methods Forum on July 21, 2021, and at University of California Davis's Quantitative Psychology Speaker Series on October 7, 2021.

The authors would like to thank Thomas R. Belin for his helpful feedback and insight.

References

Ammerman BA, Jacobucci R, & McCloskey MS (2018). Using exploratory data mining to identify important correlates of nonsuicidal self-injury frequency. Psychology of Violence, 8(4), 515–525. 10.1037/vio0000146 [PubMed: 30393574]

Arbuckle JL (1996). Full information estimation in the presence of incomplete data. In Marcoulides GA & Schumacker RE (Eds.), Advanced structural equation modeling (pp. 243–277). Erlbaum, Inc.

Austin PC, Lee DS, Ko DT, & White IR (2019). Effect of variable selection strategy on the performance of prognostic models when using multiple imputation. Circulation: Cardiovascular Quality and Outcomes, 12(11), e005927. 10.1161/CIRCOUTCOMES.119.005927 [PubMed: 31718298]

Bartlett JW, Seaman SR, White IR, & Carpenter JR (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. Statistical Methods in Medical Research, 24(4), 462–487. 10.1177/0962280214521348 [PubMed: 24525487]

Beale EML, & Little RJA (1975). Missing values in multivariate analysis. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 37(1), 129–145. 10.1111/j.2517-6161.1975.tb01037.x

Beesley LJ, & Taylor JMG (2021). Accounting for not-at-random missingness through imputation stacking. Statistics in Medicine, 40(27), 6118–6132. 10.1002/sim.9174 [PubMed: 34459011]

Belloni A, Chen D, Chernozhukov V, & Hansen C (2012). Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica, 80(6), 2369–2429. 10.3982/ECTA9626

Belloni A, & Chernozhukov V (2013). Least squares after model selection in high-dimensional sparse models. Bernoulli, 19(2), 521–547. 10.3150/11-BEJ410

Belloni A, Chernozhukov V, & Hansen C (2014a). High-dimensional methods and inference on structural and treatment effects. The Journal of Economic Perspectives, 28(2), 29–50. 10.1257/jep.28.2.29

Belloni A, Chernozhukov V, & Hansen C (2014b). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81(2), 608–650. 10.1093/restud/rdt044

Belloni A, Chernozhukov V, & Wei Y (2016). Post-selection inference for generalized linear models with many controls. Journal of Business & Economic Statistics, 34(4), 606–619. 10.1080/07350015.2016.1166116

Brand JPL (1999). Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets [PhD Thesis]. Erasmus University, Rotterdam.

Breiman L (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science, 16(3), 199–231. 10.1214/ss/1009213726

Brooks SP, & Gelman A (1998). General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics, 7(4), 434–455. 10.1080/10618600.1998.10474787

Carpenter JR, & Kenward MG (2013). Multiple imputation and its application. Wiley. 10.1002/9781119942283

Carpenter JR, Kenward MG, & White IR (2007). Sensitivity analysis after multiple imputation under missing at random: A weighting approach. Statistical Methods in Medical Research, 16(3), 259–275. 10.1177/0962280206075303 [PubMed: 17621471]

Chatterjee A, & Lahiri SN (2011). Bootstrapping lasso estimators. Journal of the American Statistical Association, 106(494), 608–625. 10.1198/jasa.2011.tm10159

Chen EE, & Wojcik SP (2016). A practical guide to big data research in psychology. Psychological Methods, 21(4), 458–474. 10.1037/met0000111 [PubMed: 27918178]

Chen Q, & Wang S (2013). Variable selection for multiply-imputed data with application to dioxin exposure study. Statistics in Medicine, 32(21), 3646–3659. 10.1002/sim.5783 [PubMed: 23526243]

Chen X, Lin Q, Kim S, Carbonell JG, & Xing EP (2012). Smoothing proximal gradient method for general structured sparse regression. The Annals of Applied Statistics, 6(2), 719–752. 10.1214/11-AOAS514

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, & Robins J (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68. 10.1111/ectj.12097

Chernozhukov V, Hansen C, & Spindler M (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. The American Economic Review, 105(5), 486–490. 10.1257/aer.p20151022

Cohen J, Cohen P, West SG, & Aiken LS (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Erlbaum.

Collins LM, Schafer JL, & Kam CM (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4), 330–351. 10.1037/1082-989X.6.4.330 [PubMed: 11778676]

Comulada WS, Goldbeck C, Almirol E, Gunn HJ, Ocasio MA, Fernández MI, Arnold EM, Romero-Espinoza A, Urauchi S, Ramos W, Rotheram-Borus MJ, Klausner JD, Swendeman D, & Adolescent Medicine Trials Network (ATN) CARES Team. (2021). Using machine learning to predict young people's internet health and social service information seeking. Prevention Science, 22(8), 1173–1184. 10.1007/s11121-021-01255-2 [PubMed: 33974226]

Comulada WS, Step M, Fletcher JB, Tanner AE, Dowshen NL, Arayasirikul S, Keglovitz Baker K, Zuniga J, Swendeman D, Medich M, Kao UH, Northrup A, Nieto O, Brooks RA, & Special Projects Of National Significance Social Media Initiative Study Group. (2020). Predictors of internet health information–seeking behaviors among young adults living with HIV across the United States: Longitudinal observational study. Journal of Medical Internet Research, 22(11), e18309. 10.2196/18309 [PubMed: 33136057]

Cui Z, & Gong G (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. NeuroImage, 178(1), 622–637. 10.1016/j.neuroimage.2018.06.001 [PubMed: 29870817]

Daubechies I, Defrise M, & De Mol C (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Communications on Pure and Applied Mathematics, 57(11), 1413–1457. 10.1002/cpa.20042

De Mol C, Mosci S, Traskine M, & Verri A (2009). A regularized method for selecting nested groups of relevant genes from microarray data. Journal of Computational Biology, 16(5), 677–690. 10.1089/cmb.2008.0171 [PubMed: 19432538]

Dempster AP, Laird NM, & Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B, Statistical Methodology, 39(1), 1–22. 10.1111/j.2517-6161.1977.tb01600.x

Deng Y, Chang C, Ido MS, & Long Q (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. Scientific Reports, 6(1), 21689. 10.1038/srep21689 [PubMed: 26868061]

Dobbin KK, & Simon RM (2011). Optimally splitting cases for training and testing high dimensional classifiers. BMC Medical Genomics, 4(1), 31–31. 10.1186/1755-8794-4-31 [PubMed: 21477282]

Donoho D (2017). 50 years of data science. Journal of Computational and Graphical Statistics, 26(4), 745–766. 10.1080/10618600.2017.1384734

Du J, Boss J, Han P, Beesley LJ, Goutman SA, Batterman S, & Mukherjee B (2020). Variable selection with multiply-imputed datasets: Choosing between stacked and grouped methods. arXiv https://arxiv.org/abs/2003.07398

Dumas D, Doherty M, & Organisciak P (2020). The psychology of professional and student actors: Creativity, personality, and motivation. PLoS ONE, 15(10), e0240728. 10.1371/journal.pone.0240728 [PubMed: 33091923]

Efron B, Hastie T, Johnstone I, & Tibshirani R (2004). Least angle regression. Annals of Statistics, 32(2), 407–451. 10.1214/009053604000000067

Enders CK (2010). Applied missing data analysis. Guilford Press.

Enders CK, Du H, & Keller BT (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. Psychological Methods, 25(1), 88–112. 10.1037/met0000228 [PubMed: 31259566]

Epskamp S, & Fried EI (2018). A tutorial on regularized partial correlation networks. Psychological Methods, 23(4), 617–634. 10.1037/met0000167 [PubMed: 29595293]

Erler NS, Rizopoulos D, Rosmalen J, Jaddoe VWV, Franco OH, & Lesaffre EMEH (2016). Dealing with missing covariates in epidemiologic studies: A comparison between multiple imputation and a full Bayesian approach. Statistics in Medicine, 35(17), 2955–2974. 10.1002/sim.6944 [PubMed: 27042954]

Fan J, & Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456), 1348–1360. 10.1198/016214501753382273

Feelders A (1999). Handling missing data in trees: Surrogate splits or statistical imputation? In Żytkow JM & Rauch J (Eds.), Principles of data mining and knowledge discovery (pp. 329–334). Springer.

Feng L, Hancock R, Watson C, Bogley R, Miller Z, Luisa GTM, Briggs-Gowan M, & Hoeft F (2020). Development of an abbreviated adult reading history questionnaire (ARHQ-Brief) using a machine learning approach. PsyArXiv. 10.31234/osf.io/8u5fe

Friedman J, Hastie T, Tibshirani R, Narasimhan B, Tay K, Simon N, & Qian J (2020). glmnet: Lasso and elastic-net regularized generalized linear models (Version 4.0–2) [R Package]. https://cran.r-project.org/web/packages/glmnet/glmnet.pdf

Gabrieli JDE, Ghosh SS, & Whitfield-Gabrieli S (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron, 85(1), 11–26. 10.1016/j.neuron.2014.10.047 [PubMed: 25569345]

Garcia RI, Ibrahim JG, & Zhu H (2010a). Variable selection for regression models with missing data. Statistica Sinica, 20(1), 149–165. [PubMed: 20336190]

Garcia RI, Ibrahim JG, & Zhu H (2010b). Variable selection in the Cox regression model with covariates missing at random. Biometrics, 66(1), 97–104. 10.1111/j.1541-0420.2009.01274.x [PubMed: 19459831]

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, & Rubin DB (2014). Bayesian data analysis (3rd ed.). CRC Press.

Geronimi J, & Saporta G (2017). Variable selection for multiply-imputed data with penalized generalized estimating equations. Computational Statistics & Data Analysis, 110, 103–114. 10.1016/j.csda.2017.01.001

Graham JW (2012). Missing data: Analysis and design (1st ed.). Springer. 10.1007/978-1-4614-4018-5

Greenland S, & Finkle WD (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. American Journal of Epidemiology, 142(12), 1255–1264. 10.1093/oxfordjournals.aje.a117592 [PubMed: 7503045]

Guyon I (1997). A scaling law for the validation-set training-set size ratio. AT&T Bell Laboratories. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.1337&rep=rep1&type=pdf

Hagerty MR, & Srinivasan V (1991). Comparing the predictive powers of alternative multiple regression models. Psychometrika, 56(1), 77–85. 10.1007/BF02294587

Hansen BE (2016). The risk of James–Stein and Lasso shrinkage. Econometric Reviews, 35(8–10), 1456–1470. 10.1080/07474938.2015.1092799

Harrell FE (2001). Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis. Springer. 10.1007/978-1-4757-3462-1

Harris J, Purssell E, Cornelius V, Ream E, Jones A, & Armes J (2020). Development and internal validation of a predictive risk model for anxiety after completion of treatment for early stage breast cancer. Journal of Patient-Reported Outcomes, 4(1), 103. 10.1186/s41687-020-00267-w [PubMed: 33275165]

Hastie T, Tibshirani R, & Friedman J (2009). The elements of statistical learning: Data mining, inference, and prediction. Springer Science & Business Media. 10.1007/978-0-387-84858-7

Hayati Rezvan P, Lee KJ, & Simpson JA (2015). The rise of multiple imputation: A review of the reporting and implementation of the method in medical research. BMC Medical Research Methodology, 15(1), 30. 10.1186/s12874-015-0022-1 [PubMed: 25880850]

Hayati Rezvan P, Lee KJ, & Simpson JA (2018). Sensitivity analysis within multiple imputation framework using delta-adjustment: Application to longitudinal study of Australian children. Longitudinal and Life Course Studies, 9(3), 259–278. 10.14301/llcs.v9i3.503

Hesterberg T, Choi NH, Meier L, & Fraley C (2008). Least angle and ℓ_1 penalized regression: A review. Statistics Surveys, 2, 61–93. 10.1214/08-SS035

Heymans MW, van Buuren S, Knol DL, van Mechelen W, & de Vet HCW (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. BMC Medical Research Methodology, 7(1), 33–33. 10.1186/1471-2288-7-33 [PubMed: 17629912]

Horton NJ, & Kleinman KP (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. The American Statistician, 61(1), 79–90. 10.1198/000313007X172556 [PubMed: 17401454]

Huang Y, & Montoya A (2020). Lasso and group lasso with categorical predictors: Impact of coding strategy on variable selection and prediction. PsyArXiv. 10.31234/osf.io/wc45u

Hung P, Osias E, Konda KA, Calvo GM, Reyes-Díaz EM, Vargas SK, Goldbeck C, Caceres CF, & Klausner JD (2020). High lifetime prevalence of syphilis in men who have sex with men and transgender women versus low lifetime prevalence in female sex workers in Lima, Peru. Sexually Transmitted Diseases, 47(8), 549–555. 10.1097/OLQ.0000000000001200 [PubMed: 32541611]

Ibrahim JG, Chen M-H, & Kim S (2008). Bayesian variable selection for the Cox regression model with missing covariates. Lifetime Data Analysis, 14(4), 496–520. 10.1007/s10985-008-9101-5 [PubMed: 18836829]

Ibrahim JG, Chen M-H, & Lipsitz SR (2002). Bayesian methods for generalized linear models with covariates missing at random. The Canadian Journal of Statistics, 30(1), 55–78. 10.2307/3315865

Immekus JC, Muntis F, & de Paleville DT (2019). Predictor selection using lasso to examine the association of motor proficiency, postural control, visual efficiency, and behavior with the academic skills of elementary school-aged children. Journal of Motor Learning and Development, 8(1), 126–144. 10.1123/jmld.2018-0023

Jackman S (2000). Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo. American Journal of Political Science, 44(2), 375–404. 10.2307/2669318

James G, Witten D, Hastie T, & Tibshirani R (2013). An introduction to statistical learning: With applications in R. Springer. 10.1007/978-1-4614-7138-7

Johnson BA (2008). Variable selection in semiparametric linear regression with censored data. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 70(2), 351–370. 10.1111/j.1467-9868.2008.00639.x

Johnson BA, Lin DY, & Zeng D (2008). Penalized estimating functions and variable selection in semiparametric regression models. Journal of the American Statistical Association, 103(482), 672–680. 10.1198/016214508000000184 [PubMed: 20376193]

Johnson M, & Sinharay S (2011). Remarks from the new editors. Journal of Educational and Behavioral Statistics, 36(1), 3–5. 10.3102/1076998610387267

Keller BT, & Enders CK (2019). Blimp user's guide (Version 2.1) [Computer Software]. www.appliedmissingdata.com/multilevel-imputation.html

Kim S, Belin TR, & Sugar CA (2018). Multiple imputation with non-additively related variables: Joint-modeling and approximations. Statistical Methods in Medical Research, 27(6), 1683–1694. 10.1177/0962280216667763 [PubMed: 27647811]

Kim S, Sugar CA, & Belin TR (2015). Evaluating model-based imputation methods for missing covariates in regression models with interactions. Statistics in Medicine, 34(11), 1876–1888. 10.1002/sim.6435 [PubMed: 25630757]

Kroenke K, Spitzer RL, & Williams JBW (2001). The PHQ-9: Validity of a brief depression severity measure. Journal of General Internal Medicine, 16(9), 606–613. 10.1046/j.1525-1497.2001.016009606.x [PubMed: 11556941]

Kyung M, Gill J, Ghosh M, & Casella G (2010). Penalized regression, standard errors, and Bayesian lassos. Bayesian Analysis, 5(2), 369–411.

Lachenbruch PA (2011). Variable selection when missing values are present: A case study. Statistical Methods in Medical Research, 20(4), 429–444. 10.1177/0962280209358003 [PubMed: 20442196]

Leacy FP, Floyd S, Yates TA, & White IR (2017). Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta adjustment: Application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. American Journal of Epidemiology, 185(4), 304–315. 10.1093/aje/kww107 [PubMed: 28073767]

Li L, Shen C, Li X, & Robins JM (2013). On weighting approaches for missing data. Statistical Methods in Medical Research, 22(1), 14–30. 10.1177/0962280211403597 [PubMed: 21705435]

Liang X, & Jacobucci R (2020). Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic net. Structural Equation Modeling, 27(5), 722–734. 10.1080/10705511.2019.1693273

Little RJA, & Rubin DB (2019). Statistical analysis with missing data (3rd ed.). Wiley-Interscience.

Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, Frangakis C, Hogan JW, Molenberghs G, Murphy SA, Neaton JD, Rotnitzky A, Scharfstein D, Shih WJ, Siegel JP, & Stern H (2012). The prevention and treatment of missing data in clinical trials. The New England Journal of Medicine, 367(14), 1355–1360. 10.1056/NEJMsr1203730 [PubMed: 23034025]

Liu Y, Wang Y, Feng Y, & Wall MM (2016). Variable selection and prediction with incomplete high-dimensional data. The Annals of Applied Statistics, 10(1), 418–450. 10.1214/15-AOAS899 [PubMed: 27213023]

Lo A, Chernoff H, Zheng T, & Lo S-H (2015). Why significant variables aren't automatically good predictors. Proceedings of the National Academy of Sciences of the United States of America, 112(45), 13892–13897. 10.1073/pnas.1518285112 [PubMed: 26504198]

Lockhart R, Taylor J, Tibshirani RJ, & Tibshirani R (2014). A significance test for the lasso. Annals of Statistics, 42(2), 413–468. [PubMed: 25574062]

Long Q, & Johnson BA (2015). Variable selection in the presence of missing data: Resampling and imputation. Biostatistics, 16(3), 596–610. 10.1093/biostatistics/kxv003 [PubMed: 25694614]

Ludtke O, Robitzsch A, & West SG (2020). Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. Psychological Methods, 25(2), 157–181. 10.1037/met0000233 [PubMed: 31478719]

Mackinnon A (2010). The use and reporting of multiple imputation in medical research - a review. Journal of Internal Medicine, 268(6), 586–593. 10.1111/j.1365-2796.2010.02274.x [PubMed: 20831627]

Makalic E, & Schmidt DF (2016). High-dimensional Bayesian regularised regression with the bayesreg package. arXiv https://arxiv.org/abs/1611.06649

Marino M, Buxton OM, & Li Y (2017). Covariate selection for multilevel models with missing data. Stat (International Statistical Institute), 6(1), 31–46. 10.1002/sta4.133 [PubMed: 28239457]

Masconi KL, Matsha TE, Erasmus RT, & Kengne AP (2015). Effects of different missing data imputation techniques on the performance of undiagnosed diabetes risk prediction models in a mixed-ancestry population of South Africa. PLoS ONE, 10(9), e0139210. 10.1371/journal.pone.0139210 [PubMed: 26406594]

McNeish DM (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. Multivariate Behavioral Research, 50(5), 471–484. 10.1080/00273171.2015.1036965 [PubMed: 26610247]

Meinshausen N, & Bühlmann P (2010). Stability selection. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 72(4), 417–473. 10.1111/j.1467-9868.2010.00740.x

Mukhopadhyay S, & Wang K (2020). Breiman's "two cultures" revisited and reconciled. arXiv. https://arxiv.org/abs/2005.13596

Musoro JZ, Zwinderman AH, Puhan MA, ter Riet G, & Geskus RB (2014). Validation of prediction models based on lasso regression with multiply imputed data. BMC Medical Research Methodology, 14(1), 116–116. 10.1186/1471-2288-14-116 [PubMed: 25323009]

Muthén LK, & Muthén BO (1998–2017). Mplus user's guide. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

Nam SM, Peterson TA, Butte AJ, Seo KY, & Han HW (2020). Explanatory model of dry eye disease using health and nutrition examinations: Machine learning and network-based factor analysis from a national survey. JMIR Medical Informatics, 8(2), e16153. 10.2196/16153 [PubMed: 32130150]

Nesterov Y (2005). Smooth minimization of non-smooth functions. Mathematical Programming, 103(1), 127–152. 10.1007/s10107-004-0552-5

Park T, & Casella G (2008). The Bayesian lasso. Journal of the American Statistical Association, 103(482), 681–686. 10.1198/016214508000000337

Parker R (2010). Missing data problems in machine learning. VDM Verlag.

Pelham WE III, Petras H, & Pardini DA (2020). Can machine learning improve screening for targeted delinquency prevention programs? Prevention Science, 21(2), 158–170. 10.1007/s11121-019-01040-2 [PubMed: 31696355]

Poulos J, & Valle R (2018). Missing data imputation for supervised learning. Applied Artificial Intelligence, 32(2), 186–196. 10.1080/08839514.2018.1448143

Quartagno M, & Carpenter JR (2019). Multiple imputation for discrete data: Evaluation of the joint latent normal model. Biometrical Journal. [Biometrische Zeitschrift], 61(4), 1003–1019. 10.1002/bimj.201800222 [PubMed: 30868652]

Quartagno M, & Carpenter JR (2020). Package 'jomo' [Package]. https://cran.r-project.org/web/packages/jomo/jomo.pdf

R Core Team. (2018). R: A language and environment for statistical computing. (Version 3.1.2) [Computer software]. http://www.R-project.org/

Raghunathan TE, Lepkowski JM, Hoewyk JV, & Solenberger P (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology, 27, 85–95.

Rodgers DM, Jacobucci R, & Grimm KJ (2021). A multiple imputation approach for handling missing data in classification and regression trees. Journal of Behavioral Data Science, 1(1), 127–153. 10.35566/jbds/v1n1/p6 [PubMed: 35281484]

Rubin DB (1987). Multiple imputation for nonresponse in surveys (1st ed.). Wiley. 10.1002/9780470316696

Rubin DB (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, 91(434), 473–489. 10.1080/01621459.1996.10476908

Ryali S, Chen T, Supekar K, & Menon V (2012). Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. NeuroImage, 59(4), 3852–3861, 10.1016/j.neuroimage.2011.11.054 [PubMed: 22155039]

Sabbe N, Thas O, & Ottoy JP (2013). EMLasso: Logistic lasso with missing data. Statistics in Medicine, 32(18), 3143–3157. 10.1002/sim.5760 [PubMed: 23440969]

Schafer JL (1997). Analysis of incomplete multivariate data. Chapman and Hall. 10.1201/9781439821862

Schafer JL (1999). Multiple imputation: A primer. Statistical Methods in Medical Research, 8(1), 3–15. 10.1177/096228029900800102 [PubMed: 10347857]

Schafer JL, & Olsen MK (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research, 33(4), 545–571. 10.1207/s15327906mbr3304_5 [PubMed: 26753828]

Seaman SR, & White IR (2013). Review of inverse probability weighting for dealing with missing data. Statistical Methods in Medical Research, 22(3), 278–295. 10.1177/0962280210395740 [PubMed: 21220355]

Shmueli G (2010). To explain or to predict? Statistical Science, 25(3), 289–310. 10.1214/10-STS330

Simon N, Friedman J, Hastie T, & Tibshirani R (2013). A sparse-group lasso. Journal of Computational and Graphical Statistics, 22(2), 231–245. 10.1080/10618600.2012.681250

Smith EN, Young MD, & Crum AJ (2019). Stress, mindsets, and success in Navy SEALs special warfare training. Frontiers in Psychology, 10, 2962. 10.3389/fpsyg.2019.02962 [PubMed: 32010023]

StataCorp. (2019). Stata statistical software: Release 16. Stata Corp LLC.

Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, & Carpenter JR (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ(338), b2393. 10.1136/bmj.b2393 [PubMed: 19564179]

Swendeman D, Arnold EM, Harris D, Fournier J, Comulada WS, Reback C, Koussa M, Ocasio M, Lee S-J, Kozina L, Fernández MI, Rotheram MJ, & Adolescent Medicine Trials Network (ATN) CARES Team. (2019). Text-messaging, online peer support group, and coaching strategies to optimize the HIV prevention continuum for youth: Protocol for a randomized controlled trial. JMIR Research Protocols, 8(8), e11165. 10.2196/11165 [PubMed: 31400109]

Tanner MA, & Wong WH (1987). The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association, 82(398), 528–540. 10.1080/01621459.1987.10478458

Thao LTP, & Geskus R (2019). A comparison of model selection methods for prediction in the presence of multiply imputed data. Biometrical Journal. [Biometrische Zeitschrift], 61(2), 343–356. 10.1002/bimj.201700232 [PubMed: 30353591]

Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 58(1), 267–288. 10.1111/j.2517-6161.1996.tb02080.x

Tompsett DM, Leacy F, Moreno-Betancur M, Heron J, & White IR (2018). On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. Statistics in Medicine, 37(15), 2338–2353. 10.1002/sim.7643 [PubMed: 29611205]

Tompsett D, Sutton S, Seaman SR, & White IR (2020). A general method for elicitation, imputation, and sensitivity analysis for incomplete repeated binary data. Statistics in Medicine, 39(22), 2921–2935. 10.1002/sim.8584 [PubMed: 32677726]

Twala B, Jones M, & Hand DJ (2008). Good methods for coping with missing data in decision trees. Pattern Recognition Letters, 29(7), 950–956. 10.1016/j.patrec.2008.01.010

van Buuren S (2018). Flexible imputation of missing data (2nd ed.). CRC Press.

van Buuren S, Boshuizen HC, & Knook DL (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine, 18(6), 681–694. 10.1002/(SICI)1097-0258(19990330)18:6<681::AID-SIM71>3.0.CO;2-R [PubMed: 10204197]

van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, & Rubin DB (2006). Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation, 76(12), 1049–1064. 10.1080/10629360600810434

Vergouwe Y, Royston P, Moons KGM, & Altman DG (2010). Development and validation of a prediction model with missing predictor data: A practical approach. Journal of Clinical Epidemiology, 63(2), 205–214. 10.1016/j.jclinepi.2009.03.017 [PubMed: 19596181]

von Hippel PT (2018). How many imputations do you need? A two-stage calculation using a quadratic rule. Sociological Methods & Research, 49(3), 699–718. 10.1177/0049124117747303

Waldmann P, Mészáros G, Gredler B, Fuerst C, & Sölkner J (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. Frontiers in Genetics, 4, 270. 10.3389/fgene.2013.00270 [PubMed: 24363662]

Wan Y, Datta S, Conklin DJ, & Kong M (2015). Variable selection models based on multiple imputation with an application for predicting median effective dose and maximum effect. Journal of Statistical Computation and Simulation, 85(9), 1902–1916. 10.1080/00949655.2014.907801 [PubMed: 26412909]

Wang S, Nan B, Rosset S, & Zhu J (2011). Random lasso. The Annals of Applied Statistics, 5(1), 468–485. 10.1214/10-AOAS377 [PubMed: 22997542]

Ware JH, Harrington D, Hunter DJ, & D'Agostino RB Sr. (2012). Missing data. The New England Journal of Medicine, 367(14), 1353–1354. 10.1056/NEJMsm1210043

Wasserman L, & Roeder K (2009). High dimensional variable selection. Annals of Statistics, 37(5A), 2178–2201. 10.1214/08-AOS646 [PubMed: 19784398]

White IR, Royston P, & Wood AM (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine, 30(4), 377–399. 10.1002/sim.4067 [PubMed: 21225900]

Wolfson J (2011). EEBoost: A general method for prediction and variable selection based on estimating equations. Journal of the American Statistical Association, 106(493), 296–305. 10.1198/jasa.2011.tm10098

Wood AM, White IR, & Royston P (2008). How should variable selection be performed with multiply imputed data? Statistics in Medicine, 27(17), 3227–3246. 10.1002/sim.3177 [PubMed: 18203127]

Wu TT, & Lange K (2008). Coordinate descent algorithms for lasso penalized regression. The Annals of Applied Statistics, 2(1), 224–244. 10.1214/07-AOAS147

Yang X, Belin TR, & Boscardin WJ (2005). Imputation and variable selection in linear regression models with missing covariates. Biometrics, 61(2), 498–506. 10.1111/j.1541-0420.2005.00317.x [PubMed: 16011697]

Yarkoni T, & Westfall J (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science, 12(6), 1100–1122. 10.1177/1745691617693393 [PubMed: 28841086]

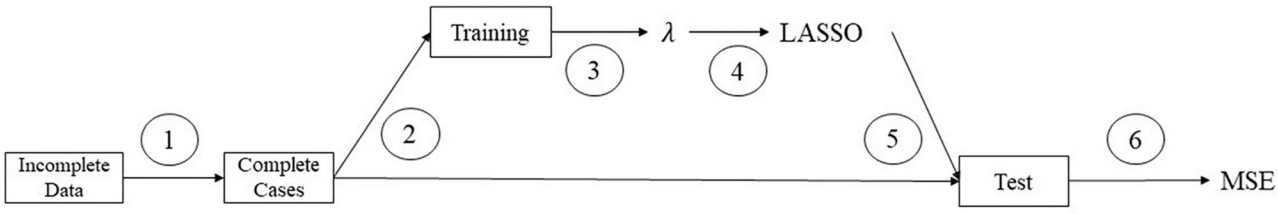
Yuan M, & Lin Y (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 68(1), 49–67. 10.1111/j.1467-9868.2005.00532.x

Zhao Y, & Long Q (2016). Multiple imputation in the presence of high-dimensional data. Statistical Methods in Medical Research, 25(5), 2021–2035. 10.1177/0962280213511027 [PubMed: 24275026]

Zhao Y, & Long Q (2017). Variable selection in the presence of missing data: Imputation-based methods. Wiley Interdisciplinary Reviews: Computational Statistics, 9(5), e1402. 10.1002/wics.1402 [PubMed: 29085552]

Zou H, & Hastie T (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 67(2), 301–320. 10.1111/j.1467-9868.2005.00503.x

Zou H, Hastie T, & Tibshirani R (2007). On the "degrees of freedom" of the lasso. Annals of Statistics, 35(5), 2173–2192. 10.1214/009053607000000127



1 Listwise deletion: drop participants with missing data on any variables in planned analysis

2 Split participants with complete data into training and test sets

3 Conduct *k*-fold cross-validation in training set to obtain optimal λ value

4 Create training model (i.e., estimate a LASSO using the training set) using optimal λ

5 Fit training model to test set

6 Calculate a model performance measure (e.g., MSE)

Figure 1. LASSO Procedure for Complete Data

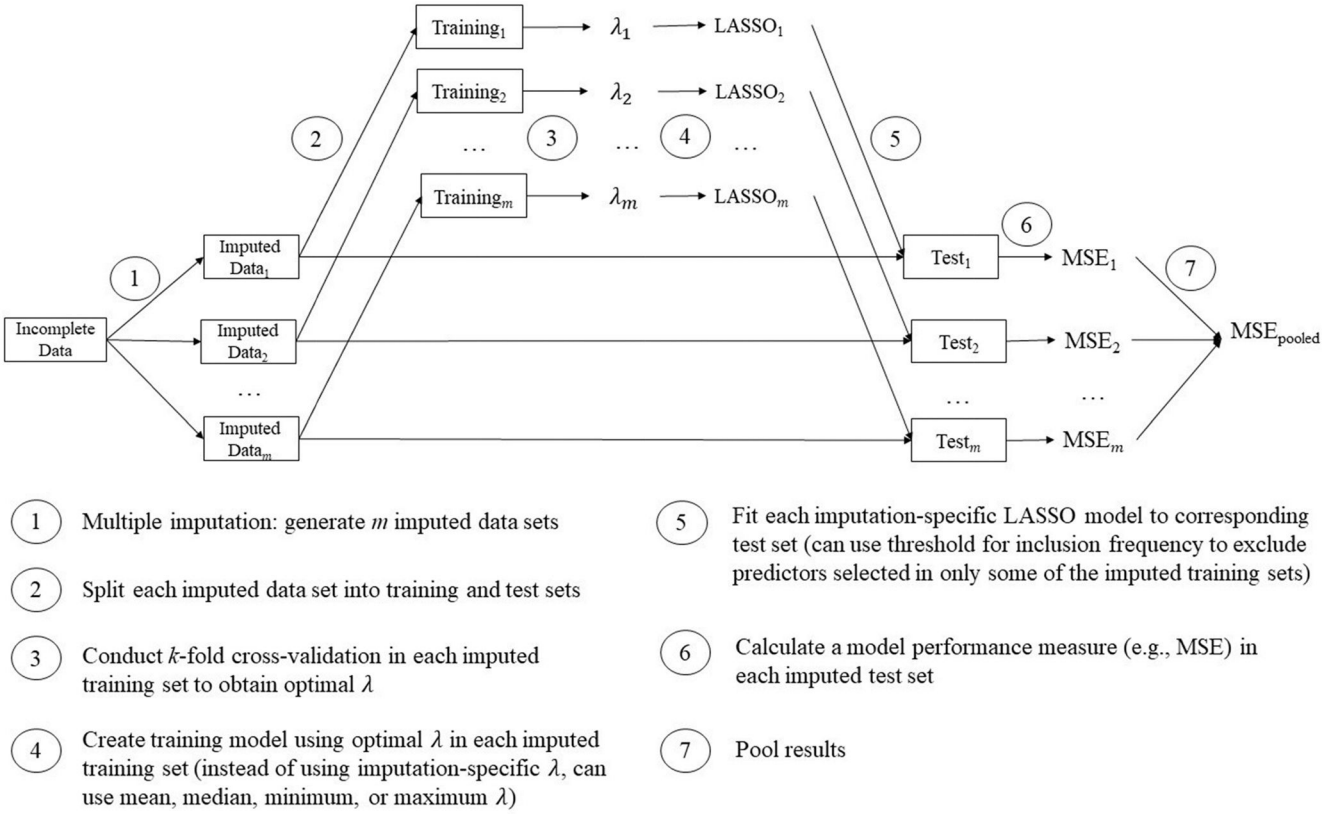


Figure 2. LASSO Procedure for Separate Approach

Note. The subscripts index the m imputed data sets.

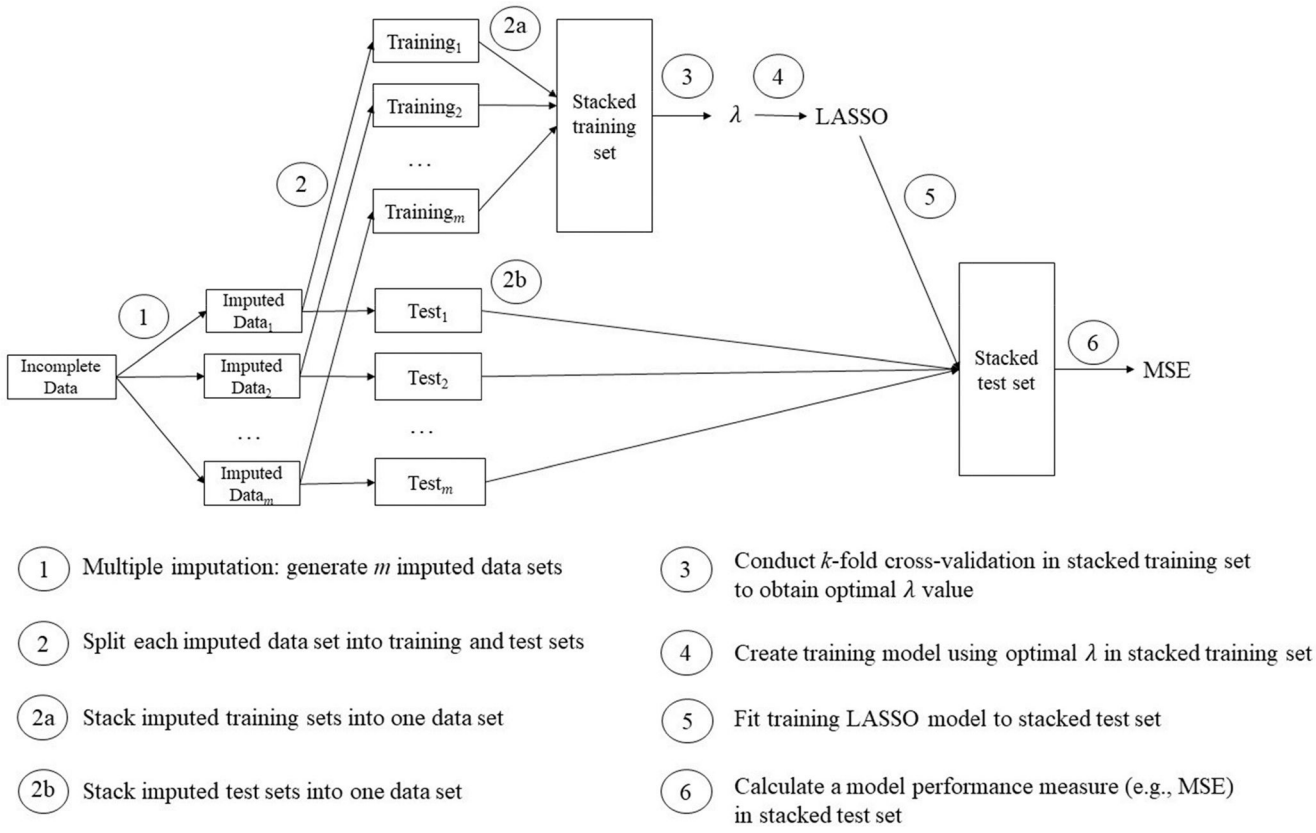


Figure 3. LASSO Procedure for Stacked Approach

Note. The subscripts index the m imputed data sets.

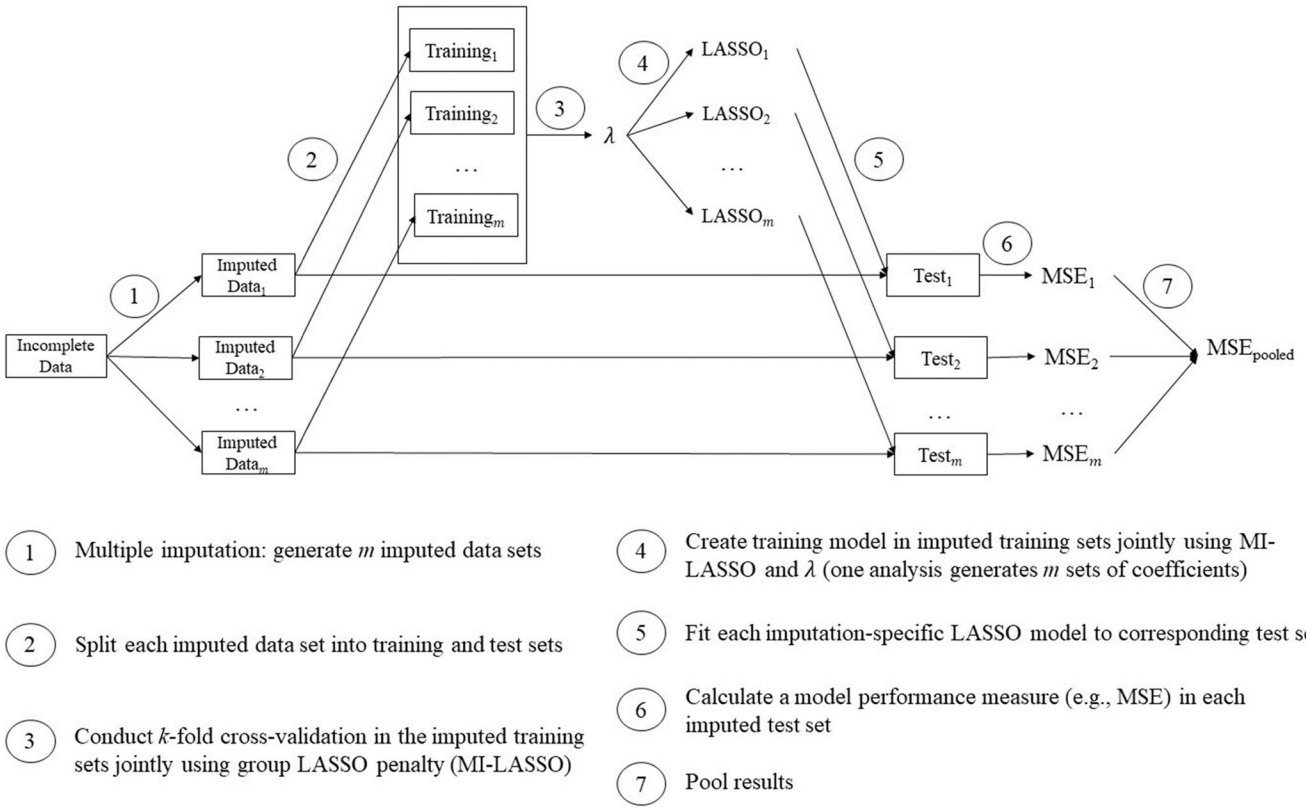


Figure 4. LASSO Procedure for MI-LASSO

Note. The subscripts index the *m* imputed data sets.

Table 1

Descriptive Statistics of Variables Entered in LASSO

Variable	*M (SD)*	*N* missing (%)
Outcome variable		
PHQ-9 depression	7.05 (5.89)	50 (3.36)
Continuous variables		
Age	20.89 (2.15)	0 (0.00)
AUDIT-C	3.01 (2.90)	15(1.01)
Count of unique drugs ever used	3.11 (2.86)	3 (0.20)
Rumination on something bad	0.78 (1.01)	11 (0.74)
SF-12 calm and peaceful	3.03 (1.40)	22(1.48)
SF-12 sad and blue	2.03 (1.39)	21 (1.41)
SF-12 emotional problems	1.69 (1.63)	27 (1.82)
SF-12 energy	3.21 (1.43)	19 (1.28)
GAD-7 anxiety	6.48 (5.52)	23 (1.55)
SR of mental health	7.02 (2.38)	3 (0.20)
SR of physical health	7.81 (1.90)	2(0.13)
SR of living situation	7.27 (2.54)	2(0.13)
SR of ability to live drug free	7.20 (3.04)	6 (0.40)
SR of social network	7.28 (2.70)	3 (0.20)
SR of sexual relationships	7.09 (2.90)	182(12.25)
Social help	7.17 (3.13)	2(0.13)
Emotional support	7.18 (3.16)	3 (0.20)
Ability to make new friends	7.88 (2.80)	4 (0.27)
Frequency of social media use	4.27 (1.56)	9 (0.61)
Frequency of dating app use	1.68 (2.05)	14 (0.94)
Binary variables		
Los Angeles	0.56 (0.50)	0 (0.00)
Black/African American	0.51 (0.50)	0 (0.00)
Latinx	0.24 (0.43)	0 (0.00)
White	0.18 (0.39)	0 (0.00)
Other race/ethnicity	0.07 (0.25)	0 (0.00)
Female at birth	0.19 (0.39)	0 (0.00)
Cisgender	0.87 (0.34)	0 (0.00)
Heterosexual	0.27 (0.44)	0 (0.00)
Employed	0.71 (0.45)	31 (2.09)
Income below poverty line	0.71 (0.45)	10 (0.67)
Has health insurance	0.80 (0.40)	122 (8.21)
Has health care provider	0.69 (0.46)	6 (0.40)
Medical utilization	0.65 (0.48)	9 (0.61)
Received ER/urgent care	0.30 (0.46)	3 (0.20)
Participated in substance abuse program	0.20 (0.40)	0 (0.00)

Variable	*M (SD)*	*N* missing (%)
Participated in HIV prevention program	0.21 (0.41)	4 (0.27)
Ever homeless	0.49 (0.50)	0 (0.00)
Ever incarcerated	0.25 (0.43)	6 (0.40)
Experienced partner violence	0.37 (0.48)	44 (2.96)
Exchanged sex for money	0.25 (0.43)	8 (0.54)
Attempted suicide	0.33 (0.47)	37 (2.49)
Hospitalized for mental health problems	0.30 (0.46)	0 (0.00)
Sexually abused	0.30 (0.46)	29 (1.95)
Had sex with someone 5+ years older before age 16	0.31 (0.46)	18 (1.21)
Ever been robbed	0.31 (0.46)	13 (0.87)
Seen serious injury or death	0.49 (0.50)	11 (0.74)
Family member was murdered	0.42 (0.49)	13 (0.87)
Used drugs during last sexual encounter	0.43 (0.50)	7 (0.47)
Ever smoked	0.45 (0.50)	45 (3.03)

Note. PHQ-9 = 9-item Patient Health Questionnaire scale; GAD-7 = 7-item Generalized Anxiety Disorder scale; AUDIT-C = scale score from Alcohol Use Disorders Identification Test; SF-12 = Item from 12-item Short Form health survey; SR = self-rating.

Table 2

Coefficient Estimates and Model Fit for Four Approaches

Variable	Listwise	Separate (IF%)	Stacked	MI-LASSO
Intercept	5.89	5.58 (100)	5.71	5.53
Age	0	0(0)	0.01	0
AUDIT-C	0	0 (0)	−0.01	0
Count of unique drugs used	0	0 (0)	−0.04	0
Rumination on something bad	0.39	0.37 (100)	0.45	0.37
SF-12 calm and peaceful	−0.21	−0.24 (100)	−0.26	−0.24
SF-12 sad and blue	0.44	0.34 (100)	0.35	0.34
SF-12 emotional problems	0.60	0.42 (100)	0.47	0.41
SF-12 energy	−0.40	−0.38 (100)	− 0.41	−0.38
GAD-7 anxiety	0.47	0.53 (100)	0.52	0.53
SR of mental health	−0.15	−0.20 (100)	−0.21	−0.20
SR of physical health	−0.12	−0.05 (100)	−0.07	−0.05
SR of living situation	0	−0.01 (78)	−0.04	−0.01
SR of ability to live drug free	0	0(0)	0.00	0
SR of social network	0	−0.00 (2)	−0.01	0
SR of sexual relationships	0	0 (0)	0.02	0
Social help	0	−0.01 (82)	−0.05	−0.01
Emotional support	0	0 (0)	0.06	0
Ability to make new friends	0	−0.01 (74)	−0.03	−0.00
Frequency of social media use	0	0.00 (2)	0.04	0
Frequency of dating app use	0	0.00 (28)	0.05	0
Los Angeles	0.25	0.71 (100)	0.87	0.70
Black/African American	−0.02	−0.10(100)	−0.14	−0.09
Latinx	0	0 (0)	0	0
White	0	0.00 (12)	0.31	0
Other race/ethnicity	0	0.00 (2)	0.16	0
Female at birth	0	0 (0)	−0.14	0
Cisgender	−0.58	−0.42 (100)	−0.59	− 0.41

Variable	Listwise	Separate (IF%)	Stacked	MI-LASSO
Heterosexual	0	0(0)	0.01	0
Employed	−0.30	−0.25 (100)	−0.50	−0.24
Income below poverty line	0	0 (0)	0.00	0
Has health insurance	0	−0.01 (16)	−0.17	0
Has health care provider	0	0 (0)	−0.06	0
Medical utilization	0.38	0.40 (100)	0.66	0.39
Received ER/urgent care	0	0(0)	−0.22	0
Participated in substance abuse program	0	−0.04 (66)	− 0.43	−0.00
Participated in HIV prevention program	0	0 (0)	−0.04	0
Ever homeless	0	0 (0)	0	0
Ever incarcerated	0	0 (0)	0.09	0
Experienced partner violence	0	0 (0)	−0.04	0
Exchanged sex for money	0	0 (0)	0.01	0
Attempted suicide	0.32	0.58 (100)	0.70	0.58
Hospitalized for mental health problems	0	0(0)	−0.02	0
Sexually abused	0.42	0.22 (100)	0.38	0.22
Sex w/ person 5+ years older before 16	0	0 (0)	−0.08	0
Ever been robbed	0.17	0.12 (98)	0.28	0.12
Seen serious injury or death	0	0 (0)	−0.19	0
Family member was murdered	0	0.01 (40)	0.39	0
Used drugs during last sexual encounter	0	−0.00 (6)	−0.21	0
Ever smoked	0	0.01 (46)	0.30	0
# of predictors selected	16	29	47	20
MSE training	10.085	10.677	10.456	10.702
MSE test	12.579	11.367	11.647	11.371
Lambda (k)	0.143	0.109	0.004	27.542

Note. IF = inclusion frequency for separate approach, percentage of imputed data sets that selected that predictor; PHQ-9 = 9-item Patient Health Questionnaire scale; GAD-7 = 7-item Generalized Anxiety Disorder scale; AUDIT-C = scale score from Alcohol Use Disorders Identification Test; SF-12 = Item from 12-item Short Form health survey; SR = self-rating; 0 indicates predictor was not selected whereas 0.00 indicates predictor was selected but the absolute value of the coefficient is less than .01.

Table 3

Coefficient Estimates From Imputation-Specific LASSO Models for Four Predictors Using the Separate Approach

Imputation	GAD-7 anxiety	Ability to make friends	Ever smoked	Age
1	0.532	0	0.017	0
2	0.527	−0.006	0	0
3	0.524	0	0.092	0
4	0.526	−0.006	0	0
5	0.514	−0.008	0	0
6	0.534	−0.001	0	0
7	0.529	−0.022	0.03	0
8	0.525	−0.013	0.026	0
9	0.516	−0.001	0	0
10	0.519	0	0.009	0
11	0.534	−0.013	0.019	0
12	0.528	−0.009	0	0
13	0.526	−0.009	0.024	0
14	0.526	−0.012	0	0
15	0.525	−0.011	0.075	0
16	0.529	−0.011	0	0
17	0.528	0	0	0
18	0.533	0	0.017	0
19	0.527	−0.004	0.034	0
20	0.515	0	0.070	0
21	0.522	0	0.009	0
22	0.521	−0.006	0.008	0
23	0.529	−0.015	0.002	0
24	0.528	−0.015	0.030	0
25	0.524	0	0	0
26	0.528	−0.005	0	0
27	0.516	0	0	0
28	0.525	−0.009	0	0
29	0.527	−0.008	0.010	0
30	0.520	−0.012	0	0
31	0.530	−0.003	0	0
32	0.538	0	0.047	0
33	0.530	0	0	0
34	0.530	−0.004	0.041	0
35	0.528	−0.003	0.008	0
36	0.533	−0.020	0	0
37	0.532	−0.008	0.078	0
38	0.529	−0.005	0	0

Imputation	GAD-7 anxiety	Ability to make friends	Ever smoked	Age
39	0.526	0	0	0
40	0.532	0	0	0
41	0.526	−0.002	0	0
42	0.526	−0.003	0	0
43	0.533	−0.014	0	0
44	0.529	−0.015	0	0
45	0.530	0	0	0
46	0.521	0	0.093	0
47	0.529	−0.005	0	0
48	0.526	0	0	0
49	0.525	−0.003	0.002	0
50	0.531	−0.018	0.009	0
Average	0.527	−0.006	0.015	0
Inclusion frequency	1.0	.74	.46	0
Coefficient in final model when				
π = 1/m	0.527	−0.006	0.015	0
π = .5	0.527	−0.006	0	0
π = 1	0.527	0	0	0

Note. Inclusion frequency = proportion of imputed data sets that selected that predictor. GAD-7 = 7-item Generalized Anxiety Disorder scale; p = threshold for inclusion in the final model.