How to Apply Variable Selection Machine Learning Algorithms with Multiply Imputed Data:

A Missing Discussion

THIS MANUSCRIPT IS UNDER REVIEW

PLEASE DO NOT COPY OR CITE WITHOUT AUTHOR’S PERMISSION

Heather J. Gunn*a*, Panteha Hayati Rezvan*a*, M. Isabel Fernández*b*, & W. Scott Comulada*a*

*a* University of California, Los Angeles

*b* Nova Southeastern University

**Author Note**

Heather J. Gunn, Panteha Hayati Rezvan, and W. Scott Comulada, Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles; M. Isabel Fernández, College of Osteopathic Medicine, Nova Southeastern University.

The first and second authors were supported by a grant from the National Institute of Mental Health (T32MH109205). The last author was also supported by a grant from the National Institute of Mental Health (P30MH058107). The data come from an NICHD-funded U19 Cooperative Agreement for the Adolescent 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; IRB #16-001674-AM-00006; Swendeman et al., 2019). The principal investigator of study protocol 149 can be contacted to access the data for this tutorial (dswendeman@mednet.ucla.edu). The analyses described in this tutorial have not been presented elsewhere.

The authors declare they have no conflicts of interest.

Correspondence concerning this article should be addressed to Heather J. Gunn, 10920 Wilshire Boulevard, STE 350, Los Angeles, CA, 90024. E-mail: HGunn@mednet.ucla.edu

**Abstract**

Psychological researchers often use standard linear regression to identify relevant predictors of an outcome of interest. Testing all predictors simultaneously can lead to overfitting and inflation of standard errors. Regularization methods like the LASSO reduce the risk of overfitting and increase model interpretability; however, handling missing data when using variable selection regularization methods is complicated. Typically, researchers use listwise deletion or *ad-hoc* single-imputation strategies to deal with missing data when using regularization methods, which can lead to loss of precision and substantial bias in the analyses. 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 through an applied example. We then compare the results of the three imputation LASSO approaches to the results obtained from traditional methods including listwise deletion and median substitution. Further, we discuss the implications for using each approach and recommend that researchers use multiple imputation over traditional missing data handling approaches when fitting a LASSO.

**Translational Abstract**

Standard linear regression is a widely-used model in psychological research that tests the relationships between hypothesized predictors and outcomes 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 and increase model interpretability. 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. We compare these three approaches to two traditional approaches that tend not to show the uncertainty associated with missing data accurately – listwise deletion and median substitution. We recommend that researchers use multiple imputation over traditional missing data handling approaches when fitting a LASSO. We provide R code so that applied researchers can implement these approaches in their own studies.

Keywords: LASSO; missing data; multiple imputation; regression; regularization

**Introduction**

Machine learning variable selection methods like the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) and elastic net (Zou & Hastie, 2005) have been overlooked by psychologists even though they reduce the risk of overfitting and improve model interpretability compared to the standard linear regression model (McNeish, 2015). A likely reason for this is the complication that arises when the data set contains 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 to reduce bias and increase power (Enders, 2010). However, the guidelines require modification for variable selection techniques. The goal of this manuscript is to provide a practical tutorial on how to implement different approaches for fitting a LASSO and by extension the elastic net when using multiple imputation to handle missing values.

Selecting relevant and important predictors of an outcome of interest is a challenging statistical problem that psychologists face (Hesterberg, Choi, Meier, & Fraley, 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 complexity in interpretation. Classical variable selection methods such as backward, forward, or stepwise selection improve model interpretability (Harrell, 2001). Variable selection methods identify a set of variables that are most associated with or predictive of the outcome of interest. They are primarily used in situations where the number of predictors is large and when it is not feasible to include all the relevant predictors and their interactions in the model. Classical variable selection methods typically provide an easily 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, reducing prediction accuracy, and difficulties with handling collinearity (Harrell, 2001).

Another way to improve upon standard linear regression is to use regularization techniques (also called penalization or shrinkage methods) to constrain or shrink the regression coefficients for predictors. Three common machine learning shrinkage methods are ridge regression, the LASSO, and elastic net. Ridge regression uses a penalty parameter to shrink the regression coefficients towards zero but does not fix them to exactly zero. The LASSO uses a different penalty parameter than ridge regression that also shrinks the estimated regression coefficients toward zero, but, unlike ridge, it shrinks some of the estimated coefficients to be 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 sets some coefficients to zero. The elastic net improves performance compared to 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 (e.g., Waldmann, Mészáros, Gredler, Fuerst, & Sölkner, 2013). In psychological research, for instance, where groups of predictors are highly correlated, elastic net may be a better choice when the goal is explanation rather than prediction; elastic net includes all the predictors in a highly correlated group. When the goal is prediction, LASSO may be preferred since it only selects one predictor to represent highly correlated groups. The approaches discussed in this tutorial apply to both LASSO and elastic net, but we focus on LASSO due to the increased calls for its widespread use in the behavioral sciences (Johnson & Sinharay, 2011; McNeish, 2015) and its popularity in simulation studies (Chen & Wang, 2013).

Missing data are almost inevitable in psychological research and can introduce substantial bias, reduce precision, and affect the generalizability of the results if not handled properly. Many researchers still rely on *ad-hoc* imputation approaches such as mean or median substitution or listwise deletion (or complete case analysis). The former approach overstates the precision of parameter estimates and gives biased results. The latter approach excludes cases with any incomplete values, which can greatly reduce the sample size, resulting in a loss of precision and statistical power, and also 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). Modern principled methods for handling missing data including maximum likelihood-based inference (Enders, 2010; Little & Rubin, 2002), multiple imputation (MI; Rubin, 1987), and inverse probability weighting (Li, Shen, Li, & Robins, 2013; Seaman & White, 2011) often rely on a formal plausible 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. MI has been increasingly used over the recent years as an accessible and flexible approach to address missing data (Hayati Rezvan, Lee, & Simpson, 2015; Mackinnon, 2010), and is now widely recommended by journal reviewers (Little et al., 2012; Ware, Harrington, Hunter, & D'Agostino, 2012). Similar to maximum likelihood approaches, MI includes all available data, but then imputes for missing values multiple times, so that it reflects uncertainty about the missing data, 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).

The standard implementation of machine learning variable selection strategies is to use listwise deletion and remove cases with incomplete data on any of the studied variables (e.g., Barrett & Lockhart, 2018; Comulada et al., 2020) Single imputation methods such as median substitution have also been utilized prior to performing variable selection (e.g., Pelham, Petras, & Pardini, 2020) Given the drawbacks of listwise deletion and *ad-hoc* imputation methods, several studies have considered how to combine the variable selection techniques with modern missing data handling techniques (Zhao & Long, 2017). There have been advances in performing variable selection using likelihood-based methods (e.g., EM-LASSO; Sabbe, Thas, & Ottoy, 2013) and within the Bayesian framework (Yang, Belin, & Boscardin, 2005). However, these techniques are underutilized in practice since they can be computationally intensive and difficult to implement by applied researchers.

More recently, there has been a growing body of literature focusing on combining variable selection techniques, particularly LASSO, with multiply imputed data. A complication of combining variable selection with multiple imputation is that conducting variable selection on each imputed data set results in different variables being selected across the imputed data sets. Although several studies evaluated variable selection techniques in the presence of missing data using simulation experiments (Chen & Wang, 2013; Thao & Geskus, 2019; Wan, Datta, Conklin, & Kong, 2015; Wood, White, & Royston, 2008), there are open questions on how to best aggregate the variable selection results across multiple imputed data sets to obtain an overall result, and less practical guidance is available for applied researchers on how to combine variable selection strategies with multiple imputation in empirical studies. This tutorial provides guidance for combining LASSO, as a machine learning variable selection technique, with multiple imputation, and discusses how to incorporate cross-validation and training and test data sets in the process, which has received little attention in the articles addressing how to combine variable selection methods with multiply imputed data.

The tutorial is organized as follows. First, we review the LASSO assuming there is complete data, the MI procedure for handling missing data, and three approaches for fitting a LASSO with multiply imputed data (henceforth referred to as *imputation LASSO approaches*). Then, using data from a longitudinal study of vulnerable youth in Los Angeles and New Orleans (Adolescent Medicine Trials Network (ATN) CARES study), we illustrate the three imputation LASSO approaches to identify important variables (e.g., demographics, social determinants, risk factors, protective acts) associated with recent depression severity among youth measured by the Patient Health Questionnaire (PHQ-9; Kroenke, Spitzer, & Williams, 2001). We perform the LASSO (1) on each of the imputed data sets; (2) on stacked imputed data sets using a weighting scheme; and (3) on the imputed data sets jointly using a group penalty (i.e., MI-LASSO; Chen & Wang, 2013). We then compare the variable selection results with the results obtained from fitting a LASSO using listwise deletion and the *ad-hoc* median substitution, two approaches for handling missing data currently being used by applied researchers when fitting a LASSO. Finally, we end the tutorial with a discussion and recommendations.

**Review of Statistical Methods**

**LASSO**

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

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where *n* is the total number of participants, is the raw score on the outcome for participant *i*, is the intercept, *p* is the total number of predictors, is the regression coefficient for predictor *j*, and is the raw score on predictor *j* for participant *i*. The terms in the brackets can be replaced by , the predicted outcome score based on the model. Equation 1 is also known as the residual sum of squares (RSS). An advantage of OLS estimation is that it is the best linear unbiased estimator. An estimator is unbiased if the sample estimates equals 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 get vastly different estimates. In other words, OLS estimates work the best in the sample used to estimate those coefficients, but they do not work well using a different sample from the same population (McNeish, 2015).

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

|  |  |  |
| --- | --- | --- |
|  |  | (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, (lambda), controls the impact of the shrinkage penalty on the estimation. If is zero, then the solution is identical to OLS estimates. If is infinity, then all regression coefficients are shrunk to zero. A tuning parameter in between these two extremes will fix some of the regression coefficients to be 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 level of description. Notice that the intercept is not included in the shrinkage penalty. This is because the intercept represents the expected value of the outcome when the values of the predictors are zero. Additionally, because of the shrinkage penalty, the scaling of predictors can lead to some predictors being favored over others simply because of scaling and not because of their strength in prediction. So, before a LASSO is fit to data, all predictors are typically standardized so that they have equal variances. Afterwards, standardized coefficients are converted to their unstandardized counterparts.

An advantage of the LASSO is that it decreases the risk of overfitting the model. By overfitting we mean modeling relationships specific to the sample that do not exist in the population. Random noise in a specific sample can be confused as signal due to sampling error. Further, interpreting a model with many predictors is difficult. By fixing some of the regression coefficients to zero, the LASSO increases the interpretability of the model. However, reducing the risk of overfitting comes at the cost of adding bias to the coefficients. The LASSO introduces bias by intentionally shrinking the coefficients, but it reduces the variance of those estimates from sample to sample. As stated earlier, OLS estimates are unbiased; however, they have high variance. This is known as the bias-variance tradeoff.

Unlike OLS, there is no closed-form solution for the LASSO. Thus, deriving standard errors for the coefficients estimated via LASSO are difficult and complex. Further, calculating the degrees of freedom is not straightforward (Zou, Hastie, & Tibshirani, 2007). Because of the complication with standard errors and degrees of freedom, the calculation of *p*-values is complex. While there have been developments in creating significance tests (e.g., Lockhart, Taylor, Tibshirani, & Tibshirani, 2014) this is beyond the scope of the tutorial. For this tutorial, we will be focusing on the variables that are selected rather than the significance of the coefficients of the selected variables.

To estimate the function in Equation 2, a value of needs to be provided. However, it is difficult to know which value of will produce the best prediction a priori. The goal is to choose a value that creates an interpretable model (i.e., shrinks some coefficients to zero), but not shrink coefficients so much that too much bias is added into the estimates. There are a few ways to determine which value to use. For this tutorial, we focus on *k*-fold cross-validation, a type of resampling method that improves the replicability of the model (James, Witten, Hastie, & Tibshirani, 2013). The steps to *k*-fold cross-validation are as follows:

1. Choose a candidate set of *l* lambda values.
2. Divide the data set into *k* roughly equally sized portions or folds.
3. Hold out the first fold as a validation sample.
4. With the remaining *k* – 1 folds, fit a LASSO for every single candidate value of and save the coefficients for each of these *l* models.
5. Test each of these *l* models in the validation sample separately. Record a measure of fit. A common measure of fit is the mean squared error (MSE; i.e., the mean of Equation 1).
6. Repeat steps three though five so that each fold acts as a validation sample one time. After this step is completed, there will be a total of *k* *l* measures of fit.
7. To obtain one measure of fit for each , average the *k* measures of fit for each candidate value of . Choose the value of associated with the best measure of fit (e.g., smallest MSE).
8. Finally, run the LASSO with the chosen value using the entire data set (i.e., all *k* folds) to obtain the final model.

Following these steps will optimize the value of and thus improve prediction.

Machine learning methods tend to be exploratory, so an important step is validating the final model in a holdout sample to assess the model’s generalizability (Chen & Wojcik, 2016). This can be done by splitting the data set into two samples: a training set of data and a test set of data. The ratio of the split that should be used is context dependent. Important factors to consider are the sample size, the ratio of signal-to-noise, and the complexity of the models being evaluated (Hastie, Friedman, & Tibshirani, 2001). All models and modifications to the model are analyzed in the training set. Then, once the final model or models are decided on, they are evaluated in the test set. If a model fits well in the training set but does not fit well in the test set, then the model does not have good predictive ability. In this tutorial, we standardize the process for training and testing the LASSO. Although there are some intricacies for the different imputation LASSO approaches, this is the process we recommend for fitting a LASSO:

1. Create a 70/30 split to split the data into a training and test set, respectively.
2. Conduct 10-fold cross-validation in the training set to find the optimal (i.e., follow steps 1-8 described above).
3. Fit a LASSO model in the entire training set using the optimal and record the training MSE and *R2*.
4. Fit the exact model from step 3 (identical regression coefficients) using the test data set and record the test MSE and *R2*.

So, more succinctly, *k*-fold cross-validation is used to find the best value in the training set. The training/test split is used to test the specific coefficients of the final model. By following these steps, the final model is more likely to replicate in a future sample than if we were to only fit one LASSO to the whole data set. As will be seen later, it is important in the context of multiple imputation to have both *k*-fold cross-validation and the training/test split to select a final model.

**Multiple Imputation**

MI consists of three phases: imputation, analysis, and pooling (Enders, 2010). In the first 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 the 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 for a single set of results that support an overall inference.

MI strategies can be divided into two general frameworks: joint modeling (JM; Schafer & Graham, 2002) and fully-conditional specification (FCS; van Buuren, Boshuizen, & Knook, 1999; van Buuren, Brand, Groothuis-Oudshoorn, & Rubin, 2006) both of which are widely available in statistical software. JM draws missing values simultaneously for all incomplete variables from a multivariate distribution (e.g., multivariate normal distribution). On the other hand, FCS, also known as chained equations or sequential regression, draws missing values iteratively from a specified set of univariate conditional distributions for each incomplete variable. In general, the JM imputation approach has a secure theoretical foundation through assuming a parametric model for multivariate data; however, issues may arise due to violation of the normality assumption and/or when variables are not measured on a continuous scale (e.g., categorical variables). Thus, decisions need to be made as to whether and how the continuous imputed values should be rounded off to the nearest discrete value (Galati, Seaton, Lee, Simpson, & Carlin, 2014; Lee, Galati, Simpson, & Carlin, 2012). Unlike JM imputation, FCS is very flexible in allowing an appropriate univariate regression specification for each variable. That is, it allows for appropriate scaling and modeling of univariate distributions as well as allows for the univariate models to incorporate non-linear terms and interaction effects. However, in practice FCS may have less theoretical appeal than JM because specified conditional distributions may be incompatible with any existing joint distributions (Bartlett, Seaman, White, & Carpenter, 2014; Meng, 1994). If a covariate has missing values and the outcome/substantive model includes interaction terms or nonlinear effects involving that covariate, then substantive model-compatible imputation (also known as model-based imputation; Bartlett et al., 2014), an extension of FCS, is preferred because otherwise substantial biases can occur due to mis-specification of the imputation model (Enders, Du, & Keller, 2020; Ludtke, Robitzsch, & West, 2020).

It is recommended that the imputation model includes as many variables as possible that predict variables with missing data or predict the mechanism giving rise to missing data (Collins, Schafer, & Kam, 2001; Graham, 2012). Auxiliary variables are variables that are not included in the target analysis model but are included in the imputation model to help improve the imputation process and increase power. When imputing missing covariates, it is important that the imputation model contains the outcome variable as an explanatory variable since it may carry information about the missing data in the covariates (Moons, Donders, Stijnen, & Harrell, 2006). Recent literature recommends that the number of imputations should be at least greater than the percentage of missing data in the analysis variables (White, Royston, & Wood, 2011) and some recommend 100 imputations or more to replicate standard errors (von Hippel, 2018).

**Imputation LASSO Approaches**

Before fitting a LASSO to multiply imputed data, the missing data need to be imputed using the FCS or JM approach via multivariate normal distribution. The imputation model includes all variables used in the LASSO and any potential auxiliary variables. After running the imputation model, an imputation LASSO approach needs to be selected. For this tutorial, we discuss three approaches in detail: the separate approach, the stacked approach, and the MI-LASSO.

***Separate Approach***

As described earlier, the typical process for analyzing multiply imputed data is to run the analysis separately in all imputed data sets and then pool the results using Rubin’s rules. This approach can be taken in the case of LASSO; however, pooling the results is not straightforward. Because each imputed data set slightly differs from one another due to different imputed values, we can fit the LASSO in each imputed data set and get a different selection of variables. One way to determine the final set of selected variables is to select those variables that were selected in at least imputed data sets ( denotes the number of imputations and is a threshold value, where ) (Heymans, Van Buuren, Knol, van Mechelen, & de Vet, 2007). Wood et al. (2008) applied a stepwise backward elimination strategy to each imputed data set and the final selection of variables was determined by the variables that were chosen in (a) at least one imputed data set (i.e., = 1/), (b) at least half of the imputed data sets (i.e., = .5), and (c) in all of the imputed data sets (i.e., = 1), and then combined the estimated parameters in each of the imputed data sets using Rubin’s rules to obtain the final model. Lachenbruch (2011) applied LARS and LASSO to each imputed data set and determined the overall variable selection result across the imputed data sets if the variables selected in at least half of the imputed data sets (i.e., = .5). Thao and Geskus (2019) also applied LASSO on multiply imputed data, where the final model was obtained by averaging the estimates of the variables that were selected in any imputed data set (i.e., = 1/) or in half of the imputed data sets (i.e., = .5). It is important to note that the final variable selection result can be very sensitive to the pre-specified threshold value , and there is no clear practical guidance available for researchers on how best to specify .

For each imputed data set in the training set, 10-fold cross-validation is conducted to determine the optimal . This implies that each imputed data set can have a different optimal value due to sampling error within imputation (different folds will lead to different values) and between imputation (different imputed values will lead to different values). We recommend averaging all lambda values to obtain one optimal value. This optimal value is then used to estimate a LASSO in each imputed data set in the training set separately. We then calculate the inclusion frequency of each predictor, which is equal to the number of times the predictor was selected divided by the number of imputed data sets. If the inclusion frequency is equal to or greater than the threshold, then that predictor is included in the final selection of variables. For this tutorial, we used the threshold of = .5 to make the final selection of variables. The coefficients for a selected variable differ from imputed data set to imputed data set. To obtain one overall solution, we recommend averaging the regression coefficients of the selected variables and intercept across the imputed data sets, which is similar to Rubin’s rules. This final model is fit in the training data set for all imputed data sets to obtain the training MSE. Then, the final model is fit in the test data set for all imputed data sets to obtain the test MSE.

***Stacked Approach***

An alternative approach, which does not lead to different sets of selected variables, is to fit the LASSO to the stacked set of imputed data sets using a specific weighting scheme. 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* *m* rows for *n* participants and *m* imputed data sets. The LASSO is fit to the entire stacked data set rather than to the individual imputed data sets. Applying the LASSO on a stacked imputed data set seems more attractive than applying the LASSO on each imputed data set separately since it does not require selecting the pre-specified threshold , which is an arbitrary decision. Wood et al. (2008) conducted classical variable selection using the weights (a) = 1/, (b) = (1 – ) / (where denotes the proportion of missing data across all variables), and (c) = (1 – )/ (where denotes the proportion of missing data for variable ) to each variable in the stacked imputed data set (see also Austin, Lee, Ko, & White, 2019). Wan et al. (2015) and Thao and Geskus (2019) performed machine learning variable selection (elastic net and LASSO, respectively) on the stacked imputed data set utilizing a weighting scheme where each individual (not variable) received a weight depending on their amount of missing data. Specifically, they assigned the weight , where is the number of predictor variables without missing values for participant , is the total number of predictors, and is the total number of imputed data sets. However, a drawback of this weighting system is that it does not consider whether the outcome is present or missing for the participant. So, a person with data for all predictors and the outcome will get the same weight as a person with data for all predictors and a missing outcome. For this tutorial, we propose to assign the following weight to each row for participant *i*: . Here *v* denotes the total number of variables (i.e., number of predictor variables plus one to include the outcome) and is the total number of variables without missing values for participant *i*. If a participant has complete data on all variables, then the weight for their row in one imputed data set is equal to 1/*m*. By summing across the imputed data sets, then the total weight for one participant with complete data is equal to 1. Implementing the LASSO in this case is straightforward as there is one complete data set except that weights need to be included in the analysis.

***MI-LASSO***

The MI-LASSO uses a group LASSO penalty to fit the LASSO on all imputed data sets jointly. When using the group LASSO, a set of variables can be entered as a group (Yuan & Lin, 2006). This is sometimes seen when entering an interaction and its main effects in the LASSO so that the main effects are chosen if the interaction is chosen or when a set of dummy variables that represent the same nominal variable 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 group LASSO can be applied to multiply imputed data (Chen & Wang, 2013). Like the separate approach, each imputed data set has its own set of estimated regression coefficients (not counting the intercept). Unlike the separate approach, the imputed data sets are analyzed jointly in one analysis rather than separately. Each predictor acts as its own group such that the regression coefficients associated with one predictor will either all be zero or all be nonzero.

After the MI-LASSO is fit to the training data, the result of the MI-LASSO is sets of regression coefficients that vary across the imputed data sets, just like the separate approach. But unlike the separate approach, if the coefficient for a particular variable is zero in one imputed data set, then it will be zero in all other imputed data sets due to the group LASSO penalty, resulting in consistent variable selection and no need for the arbitrary pre-specified threshold parameter, . However, the nonzero coefficients will not be identical across imputed data sets. To calculate a final model, we recommend averaging the coefficients across the imputed data sets to have one set of coefficients, just like in the separate approach. This will give one overall solution rather than multiple solutions particular to the different imputed data sets.

**Applied Data Example**

To illustrate how to implement these three imputation LASSO approaches, we conducted a secondary data analysis on a data set from an intervention study. Additionally, we compared the three imputation LASSO approaches to listwise deletion and median substitution as two traditional approaches for handling missing data that have been used in empirical studies when implementing machine learning variable selection. The Blimp 2.1 application (Keller & Enders, 2019) was used to conduct multiple imputation. R software v. 3.1.2 (R Core Team, 2018) was used to fit a LASSO for all five approaches. The complete codes can be found in the supplementary materials.

**Description of Data**

The data for the applied example come from a randomized controlled trial conducted through ATN CARES (study protocol 149). 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 description of the study design and recruitment strategy can be found in Swendeman et al. (2019). For this tutorial, we performed a secondary data analysis and used the data from the 1,486 adolescents 14 to 24 years of age (*M* = 20.89; *SD* = 2.15) who participated in the study. The adolescents varied in terms of gender identity, sexual orientation, race/ethnicity, and risk factors. The descriptive statistics of demographic characteristics and the other variables used for this applied example are presented in Table 1.

All variables used in the LASSO were self-reported. The outcome variable was a continuous scale score from the nine-item Patient Health Questionnaire (PHQ-9) that indicated a participant’s severity of depression in the last two weeks. We used 47 variables in the imputation model: 1 outcome variable, 1 nominal predictor variable with 4 categories (i.e., race/ethnicity), 25 binary predictor variables, and 20 continuous predictor variables. Some categories of included variables were collapsed due to low cell counts (e.g., categories of gender identity were collapsed into only two categories: cisgender and transgender). If there are low cell counts for the crosstabs of categorical variables, then the imputation model often has issues with converging. The variables used in the analysis were chosen based on potential theoretical relevance to depression.

It is important to note that if a predictor variable is categorical with more than two categories, there is debate as to how to best represent the categories in the LASSO model (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). This creates a singular (and noninvertible) 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. The LASSO, however, can estimate a solution with a singular design matrix, typically 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 approaches. Specifically, the four race/ethnicity categories were represented by four dummy variables. So even though there were 47 variables in the imputation model (race/ethnicity entered as one variable), there were a total of 50 variables used in the LASSO model (race/ethnicity entered as four variables) for this applied example. When discussing details related to the imputation model (e.g., missing data rates), we refer to there being 47 variables (46 predictors and 1 outcome), but when discussing details related to the LASSO (e.g., results of the LASSO), we refer to there being 50 variables (49 predictors and 1 outcome).

**Data Analytic Strategy**

The outcome variable and many of the predictor variables had missing data (see Table 1). The percentage of missing values across the 47 variables varied between 0 and 12%. In total, 837 out of 471,486 = 69,842 (1.20%) records were incomplete. Restricting statistical analysis to participants with complete observations on all the 47 variables (i.e., listwise deletion) resulted in discarding 482 (32%) participants. If participants with complete data systematically differ from those with incomplete data, the complete cases would not be representative of all the participants in the study sample; listwise deletion may produce biased results. Thus, we adopted MI using FCS to handle missing values. No auxiliary variables were included in the imputation model. We generated *m* = 50 imputed data sets to achieve a precise inclusion frequency, more power, and better precision in point estimates (von Hippel, 2018). Convergence of the Markov chain Monte Carlo (MCMC) algorithm was determined by calculating potential scale reduction factors (PSRF) for each parameter and examining trace plots of the parameters, which are the plots of estimated parameters against the MCMC iteration numbers, to evaluate mixing of the Markov chains. If all PSRF values were below the cut-off point of 1.10, then we concluded that the model converged (Gelman et al., 2014). The burn-in period and between-imputation interval were determined by the number of iterations it took for PSRF values to drop below 1.10.

For this tutorial, we compared the three imputation LASSO approaches to two common missing data handling approaches used by applied researchers: listwise deletion and median substitution. To implement listwise deletion, a participant with missing values on any variable used in the LASSO is removed from the analysis. If using median substitution, the median of each variable with missing data (mode for categorical variables) used in the LASSO is calculated. For any variables with missing data, the median of that variable replaces the missing values. Fitting the LASSO for these two approaches for handling missing data is straightforward as there is one complete data set.

Except for the MI-LASSO approach, we analyzed all LASSOs using the package glmnet (Friedman et al., 2020) to select an optimal subset of variables most related to recent depression. To determine the appropriate penalty parameter, , we specified a 10-fold cross-validation in the training data set via the cv.glmnet function to find the that minimized the cross-validated error.

The original authors of the MI-LASSO provide SAS code and an R function for using this method upon request (Chen & Wang, 2013). We modified the provided R code to calculate the MSE rather than the BIC. This was done to match the fit measure used in the other approaches and because the BIC relies on degrees of freedom, which, as stated before, are not straightforward to calculate for LASSO. Additionally, we removed the part of the code that calculates OLS estimates as this is not needed to implement their method and it prevented us from running an over-identified model for nominal variables with more than two categories. Further, there is not an automated option to conduct 10-fold cross-validation for the MI-LASSO approach so we generated the code to do so (see supplemental materials for R code).

To increase the generalizability of the results and decrease risk of overfitting, we followed the procedure described in the “LASSO” section above and randomly split the sample into a training set (70% of participants) and a test set (30% of participants). Once a final model was chosen based on the training set, it was validated in the test set to calculate the accuracy of the model. We calculated the MSE and *R2* for both the training and test sets.

A practical consideration is how to randomly split data between the training and test sets. If the data set is organized such that each row represents a participant and the order of rows is based on when the participant entered the study, then taking the first 70% of participants would not be a random split. We recommend randomizing the data set prior to imputation. There is added complication when making a split with multiply imputed data because each row for a participant needs to be in the same split. An additional complication specific to this tutorial is that the listwise deletion approach only included participants with complete data, resulting in a different sample size compared to the other four approaches. To achieve a 70/30 split for all approaches, we first divided the sample into two non-overlapping sections: one contained all participants included in the listwise deletion set and the other section contained participants who were not. For each section, we performed a random 70/30 split. The full training data set contained the 70% from the listwise section and the 70% from the non-listwise section. The full test set contained the 30% from the listwise section and the 30% from the non-listwise section.

**Results**

The MCMC algorithm converged after 800 iterations (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.

The five approaches varied in the number of variables selected. The stacked approach selected 44 of the 49 possible predictor variables whereas listwise deletion, median substitution, the separate approach, and MI-LASSO selected 26, 18, 22, and 20 variables, respectively. Seventeen variables were selected by all five approaches. Not surprisingly, many of the mental health measures such as anxiety as measured by the GAD-7 were selected as important correlates of depression for all five approaches. The coefficients of the final models for each approach are reported in Table 2. The inclusion frequencies for the separate approach varied from 0% to 100%. Eleven variables were selected in at least one but less than 50% of the imputed data sets using this approach.

The interpretation of the coefficients differs across the five approaches due to the different selected variables. This is especially seen with the intercept and the race/ethnicity variable. The intercept is expected to vary because it is a conditional effect. Specifically, it is the expected value of depression for participants who score 0 on all selected variables from the LASSO. Because the stacked approach selected age, which is uncentered, the intercept for that approach is interpreted as the expected value of depression for participants who are aged 0 (as well as 0 on all other selected predictors). This partly explains why the intercept value for this approach is noticeably different compared to the other four approaches. The interpretation of the four dummy variables that represent race/ethnicity also differs across the approaches. For the listwise deletion approach, no dummy code representing race/ethnicity was selected so race/ethnicity was not a strong correlate of depression according to this approach. The median substitution, separate, and MI-LASSO approaches selected the Black/African American dummy variable but no other race/ethnicity dummy variables. This means the comparison in the final model for these approaches is between the Black/African American participants and all other participants. The stacked approach selected the Black/African American, Latinx, White, and Other Race dummy variables. Thus, no reference group was selected, a drawback of the Stata coding strategy for categorical variables with more than two categories (Huang & Montoya, 2020).

The optimal 𝜆 values differed for the five approaches as seen in the last row of Table 2. The stacked approach had the smallest 𝜆 value, which explains why it selected more variables than the other approaches. For the separate approach, the optimal 𝜆 values varied from 0.070 to 0.147 across the imputed data sets resulting in an average value of 0.115. Listwise deletion and median substitution had similar 𝜆 values to the separate approach. The optimal 𝜆 value for the MI-LASSO approach was much larger compared to the other approaches, but roughly the same number of variables were selected compared to listwise deletion, median substitution, and the separate approach. Table 2 presents the MSEs and *R2* for both the training and test sets. The amount of increase in the MSE between the training and test sets gives an indication of the replicability of the final model in a holdout sample. The biggest difference between the training MSE and test MSE was 1.956 for the stacked approach. The smallest difference was 1.143 for the MI-LASSO approach.

**Discussion**

This article provides a tutorial on three approaches for fitting a LASSO with multiply imputed data to support machine learning-based variable selection in the presence of missing 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 articles since that time (e.g., Donoho, 2017; Mukhopadhyay & Wang, 2020), have advocated for a blended culture so that the best method for a given task is chosen, such as LASSO for variable selection. The presence of missing data, a common and practical issue in psychosocial 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 imputation methods are more straightforward and more easily implemented through available statistical software relative to machine learning methods. For example, glmnet, used to conduct LASSO in this tutorial, only permits complete case analysis and randomForest, another R package used for random forests, only imputes single values (e.g., median substitution).

A strength of this tutorial was the provision of step-by-step procedures that can be applied to freely available 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. For instance, Chen and Wang (2013) did not conduct cross-validation to obtain when using their approach. Additionally, their code for implementing LASSO is not set up to use the results to predict scores in a hold-out sample. In this tutorial, we provided clear guidance and R code on how to conduct cross-validation 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 multiple imputation and LASSO in this article, respectively), as well as commercial software that supports multiple imputation and LASSO (e.g., Stata 16). 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 using weights. 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 unavailable, BIC is the only fit measure provided, elastic net would need to be programmed, code would need to be modified if using a categorical outcome). Additionally, it is the most computationally intensive of the three procedures. In our analyses, MI-LASSO took over four hours to run compared to seconds for the separate and stacked approaches. (Most of the computation time of the MI-LASSO was due to the cross-validation to find the optimal lambda. This time can be reduced if using an optimal method to determine the optimal lambda value.)

Another deciding factor in selecting a multiple imputation approach is interpretability. The separate approach requires an arbitrary threshold to be used in selecting regression coefficient estimates that are included in averaged estimates. And it is not clear if regression coefficient estimates that the LASSO shrinks to zero should contribute to the average for selected variables, further complicating interpretation. The stacked approach avoids selection of an arbitrary threshold but, in our example, all but five of the 49 predictor variables were selected, which makes for a less interpretable model compared to the other approaches. Additionally, the stacked approach had the worst generalizability potential (i.e., biggest difference between the training set MSE and test set MSE). The increased sample size of the stacked data set may affect the value of lambda, leading to an inaccurate number of variables being selected. 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 to the stacked approach.

Multiple imputation provided slightly favorable model fit statistics (i.e., MSE and *R2*) over listwise deletion and median substitution in the applied example. In comparing test set *R2* between the five approaches, the biggest improvement was seen for MI-LASSO relative to listwise deletion (.652 versus .604). The smallest test set MSE between the five approaches was seen for MI-LASSO relative to median substitution (11.651 versus 12.441). It is difficult draw a clear winner from the three imputation LASSO approaches, but the stacked approach appears to perform the worst of the three. However, it is important to provide caution in generalizing these findings as the analyses were applied to a single data set. Simulation results from prior studies have yet to confirm a preferred approach for analyzing and pooling the results for multiply data, especially in striking a balance between improvements in performance across multiple metrics and low implementation burden.

This tutorial described in detail three different approaches for fitting a LASSO to multiply imputed data; however, we did not provide an exhaustive list of approaches. For instance, Liu, Wang, Feng, and Wall (2016) proposed a Multiple Imputation Random LASSO (MIRL) method to select variables when using multiple imputation to handle missing data. Additionally, bootstrapping can be used in combination with multiple imputation when conducting variable selection (Heymans et al., 2007; Long & Johnson, 2015; Thao & Geskus, 2019). We chose to illustrate the separate, stacked, and MI-LASSO approaches due to their ease of implementation through readily available software and/or theoretical basis.

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 psychosocial studies call for evaluations of nonlinear relationships (e.g., interaction effects) between predictors and discrete outcomes using longitudinal data. Furthermore, multiple imputation assumes that the missing data mechanism is conditional on observed variables included in the imputation process (i.e., data are missing at random (Little & Rubin, 2002)). In practice, data may be dependent on the unobserved values of the outcome variable (i.e., not missing at random), and lead to biased statistical inference.

**Concluding Remarks**

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. We recommend using multiple imputation over listwise deletion or median substitution to handle missing data when fitting a LASSO. Given the pros and cons of each imputation LASSO approach, we refrain from recommending one LASSO approach over another and provide steps for researchers to use any of the three imputation LASSO approaches. Additionally, the separate and stacked approaches can easily be implemented with elastic net instead of the LASSO.

This tutorial also provides the basis for a much-needed discussion and evaluation of best practices for an analytic framework in which to apply traditional statistical and machine learning methods to data sets with missing values. 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, 2019) and regularized partial correlation networks (Epskamp & Fried, 2018). The frameworks can also be applied to other machine learning and artificial intelligence algorithms, such as random forests, encountered in psychosocial studies.

**References**

Austin, P. C., Lee, D. S., Ko, D. T., & White, I. R. (2019). Effect of variable selection strategy on the performance of prognostic models when using multiple imputation. *Circulation-Cardiovascular Quality and Outcomes, 12*(11). doi:10.1161/CIRCOUTCOMES.119.005927

Barrett, T. S., & Lockhart, G. (2018). Efficient exploration of many variables and interactions using regularized regression. *Prevention Science, 20*(4), 575-584. doi:10.1007/s11121-018-0963-9

Bartlett, J. W., Seaman, S. R., White, I. R., & Carpenter, J. R. (2014). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research, 24*(4), 462-487. doi:10.1177/0962280214521348

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

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods, 21*(4), 458-474. doi:10.1037/met0000111

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

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330-351. doi:10.1037//1082-989X.6.4.330

Comulada, W., Step, M., Fletcher, J., Tanner, A., Dowshen, N., Arayasirikul, S., . . . 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. *Journal of Medical Internet Research, (in press)*.

Donoho, D. (2017). 50 years of data science. *Journal of Computation and Graphical Statistics, 26*(4), 745-766. doi:10.1080/10618600.2017.1384734

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY Guilford Press.

Enders, C. K., Du, H., & Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods, 25*(1), 88-112. doi:10.1037/met0000228

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods, 23*(4), 617-634. doi:10.1037/met0000167

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]. Retrieved from <https://cran.r-project.org/web/packages/glmnet/glmnet.pdf>

Galati, J. C., Seaton, K. A., Lee, K. J., Simpson, J. A., & Carlin, J. B. (2014). Rounding non-binary categorical variables following multivariate normal imputation: Evaluation of simple methods and implications for practice. *Journal of Statistical Computation and Simulation, 84*(4), 798-811. doi:10.1080/00949655.2012.727815

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Graham, J. W. (2012). *Missing data: Analysis and design* (1st ed.). New York: Springer.

Greenland, S., & Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology, 142*(12), 1255-1264. doi:10.1093/oxfordjournals.aje.a117592

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York, NY: Springer.

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer

Hayati Rezvan, P., Lee, K. J., & Simpson, J. A. (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. doi:10.1186/s12874-015-0022-1

Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least angle and ℓ1 penalized regression:

A review. *Statistics Surveys, 2*, 61-93.

Heymans, M. W., Van Buuren, S., Knol, D. L., van Mechelen, W., & de Vet, H. C. W. (2007). Variable selection under multiple imputation using the bootstrap in a prognostic study. *BMC Medical Research Methodology, 7*(1), 33-33. doi:10.1186/1471-2288-7-33

Horton, N. J., & Kleinman, K. P. (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. doi:10.1198/000313007X172556

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York: Springer.

Johnson, M., & Sinharay, S. (2011). Remarks From the New Editors. *Journal of educational and behavioral statistics, 36*(1), 3-5. doi:10.3102/1076998610387267

Keller, B. T., & Enders, C. K. (2019). Blimp User’s Manual (Version 2.1). In. Los Angeles, CA.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*(9), 606-613. doi:10.1046/j.1525-1497.2001.016009606.x

Lachenbruch, P. A. (2011). Variable selection when missing values are present: A case study. *Statistical Methods in Medical Research, 20*(4), 429-444. doi:10.1177/0962280209358003

Lee, K. J., Galati, J. C., Simpson, J. A., & Carlin, J. B. (2012). Comparison of methods for imputing ordinal data using multivariate normal imputation: A case study of non-linear effects in a large cohort study. *Statistics in Medicine, 31*(30), 4164-4174. doi:10.1002/sim.5445

Li, L., Shen, C., Li, X., & Robins, J. M. (2013). On weighting approaches for missing data. *Statistical Methods in Medical Research, 22*(1), 14-30. doi:10.1177/0962280211403597

Liang, X., & Jacobucci, R. (2019). Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic Net. *Structural Equation Modeling*, 1-13. doi:10.1080/10705511.2019.1693273

Little, R. J. A., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., . . . Stern, H. (2012). The prevention and treatment of missing data in clinical trials. *The New England Journal of Medicine, 367*(14), 1355-1360. doi:10.1056/nejmsr1203730

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). US: Wiley-Interscience.

Liu, Y., Wang, Y., Feng, Y., & Wall, M. M. (2016). Variable selection and prediction with Incomplete high-dimensional data. *The Annals of Applied Statistics, 10*(1), 418-450. doi:10.1214/15-AOAS899

Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics, 42*(2), 413-468. doi:10.1214/13-aos1175

Long, Q., & Johnson, B. A. (2015). Variable selection in the presence of missing data: resampling and imputation. *Biostatistics, 16*(3), 596-610. doi:10.1093/biostatistics/kxv003

Ludtke, O., Robitzsch, A., & West, S. G. (2020). Regression models involving nonlinear effects with missing data: A sequential modeling approach using Bayesian estimation. *Psychological Methods, 25*(2), 157-181. doi:10.1037/met0000233

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

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research, 50*(5), 471. doi:10.1080/00273171.2015.1036965

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9*(4), 538-558. doi:10.1214/ss/1177010269

Moons, K. G. M., Donders, R. A. R. T., Stijnen, T., & Harrell, F. E. (2006). Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology, 59*(10), 1092-1101. doi:10.1016/j.jclinepi.2006.01.009

Mukhopadhyay, S., & Wang, K. (2020). Breiman's "Two Cultures" Revisited and Reconciled. *arXiv*.

Pelham, W. E., Petras, H., & Pardini, D. A. (2020). Can machine learning lmprove screening for targeted delinquency prevention programs? *Prevention Science, 21*(2), 158-170. doi:10.1007/s11121-019-01040-2

R Core Team. (2018). R: A language and environment for statistical computing. (Version 3.1.2) [Software]. Vienna, Austria. Retrieved from <http://www.R-project.org/>

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (1st ed.). New York: Wiley.

Sabbe, N., Thas, O., & Ottoy, J. P. (2013). EMLasso: Logistic lasso with missing data. *Statistics in Medicine, 32*(18), 3143-3157. doi:10.1002/sim.5760

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177. doi:10.1037/1082-989X.7.2.147

Seaman, S. R., & White, I. R. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research, 22*(3), 278-295. doi:10.1177/0962280210395740

StataCorp. (2019). Stata Statistical Software: Release 16. College Station, TX: StataCorp LLC.

Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., . . . Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ, 338*, b2393. doi:10.1136/bmj.b2393

Swendeman, D., Arnold, E. M., Harris, D., Fournier, J., Comulada, W. S., Reback, C., . . . Adolescent Medicine Trials Network, C. T. (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.

Thao, L. T. P., & Geskus, R. (2019). A comparison of model selection methods for prediction in the presence of multiply imputed data. *Biometrical Journal, 61*(2), 343-356. doi:10.1002/bimj.201700232

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

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine, 18*(6), 681-694.

van Buuren, S., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation, 76*(12), 1049-1064. doi:10.1080/10629360600810434

von Hippel, P. T. (2018). How many imputations do you need? A two-stage calculation using a quadratic rule. *Sociological methods & research, 49*(3), 004912411774730-004912411774718. doi: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.

Wan, Y., Datta, S., Conklin, D. J., & 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. doi:10.1080/00949655.2014.907801

Ware, J. H., Harrington, D., Hunter, D. J., & D'Agostino, R. B. (2012). Missing data. *The New England Journal of Medicine, 367*(14), 1353-1354. doi:10.1056/NEJMsm1210043

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine, 30*(4), 377-399. doi:10.1002/sim.4067

Wood, A. M., White, I. R., & Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine, 27*(17), 3227-3246. doi:10.1002/sim.3177

Yang, X., Belin, T. R., & Boscardin, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. *Biometrics, 61*(2), 498-506. doi:10.1111/j.1541-0420.2005.00317.x

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. doi:10.1111/j.1467-9868.2005.00532.x

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

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. doi:10.1111/j.1467-9868.2005.00503.x

Zou, H., Hastie, T., & Tibshirani, R. (2007). On the “degrees of freedom” of the lasso. *The Annals of Statistics, 35*(5), 2173-2192. doi:10.1214/009053607000000127

Table 1

*Descriptive Statistics of Variables Entered in LASSO*

|  |  |  |
| --- | --- | --- |
| Variable | *M* (*SD*) | Number 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 | 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) |
| 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) |

*Notes*. 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 Five Approaches*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Listwise deletion | Median substitution | Stacked  approach (IF%) | | Separate approach | MI-LASSO |
| Intercept | 5.99 | 6.11 | 6.01 | N/A | 5.36 | 5.95 |
| Age | 0 | 0 | 0 | (0) | 0.03 | 0 |
| AUDIT-C | 0 | 0 | 0 | (0) | -0.03 | 0 |
| Count of Unique Drugs Used | 0 | 0 | 0 | (16) | 0.01 | 0 |
| Rumination on Something Bad | 0.37 | 0.38 | 0.37 | (100) | 0.41 | 0.37 |
| SF-12 Calm and Peaceful | -0.25 | -0.32 | -0.31 | (100) | -0.34 | -0.31 |
| SF-12 Sad and Blue | 0.35 | 0.24 | 0.25 | (100) | 0.27 | 0.25 |
| SF-12 Emotional Problems | 0.56 | 0.39 | 0.39 | (100) | 0.44 | 0.39 |
| SF-12 Energy | -0.45 | -0.40 | -0.44 | (100) | -0.47 | -0.43 |
| GAD-7 Anxiety | 0.49 | 0.51 | 0.52 | (100) | 0.51 | 0.52 |
| SR of Mental Health | -0.20 | -0.25 | -0.23 | (100) | -0.24 | -0.23 |
| SR of Physical Health | -0.07 | -0.04 | -0.06 | (100) | -0.06 | -0.06 |
| SR of Living Situation | 0.03 | 0 | 0 | (0) | 0 | 0 |
| SR of Ability to Live Drug Free | 0 | 0 | 0 | (4) | -0.03 | 0 |
| SR of Social Network | 0 | 0 | 0 | (0) | 0 | 0 |
| SR of Sexual Relationships | -0.03 | 0 | 0 | (22) | 0 | 0 |
| Social Help | 0 | 0 | 0 | (0) | -0.02 | 0 |
| Emotional Support | 0.06 | 0 | 0 | (16) | 0.10 | 0 |
| Ability to Make New Friends | -0.06 | -0.01 | -0.01 | (90) | -0.07 | -0.01 |
| Frequency of Social Media Use | 0 | 0 | 0 | (6) | 0.08 | 0 |
| Frequency of Dating App Use | 0.08 | 0 | 0.02 | (84) | 0.06 | 0.01 |
| Los Angeles | 0.08 | 0.47 | 0.55 | (100) | 0.66 | 0.54 |
| Black/African American | 0 | -0.21 | -0.06 | (80) | -0.22 | -0.04 |
| Latinx | 0 | 0 | 0 | (0) | -0.05 | 0 |
| White | 0 | 0 | 0 | (0) | 0.02 | 0 |
| Other Race | 0 | 0 | 0 | (28) | 0.24 | 0 |
| Female at Birth | 0 | 0 | 0 | (0) | 0.03 | 0 |
| Cisgender | -0.76 | -0.61 | -0.62 | (100) | -0.75 | -0.60 |
| Heterosexual | 0 | 0 | 0 | (16) | -0.18 | 0 |
| Employed | -0.15 | 0 | -0.03 | (60) | -0.29 | 0 |
| Income Below Poverty Line | 0.13 | 0.05 | 0.08 | (98) | 0.29 | 0.07 |
| Has Health Insurance | -0.13 | 0 | -0.05 | (60) | -0.31 | -0.01 |
| Has Health Care Provider | 0 | 0 | 0 | (4) | -0.19 | 0 |
| Medical Utilization | 0.72 | 0.64 | 0.66 | (100) | 0.87 | 0.64 |
| Received ER/Urgent Care | 0 | 0 | 0 | (0) | 0.04 | 0 |
| Participated in Substance Abuse Program | 0 | 0 | 0 | (20) | -0.47 | 0 |
| Participated in HIV Prevention Program | 0 | 0 | 0 | (0) | -0.06 | 0 |
| Ever Homeless | 0 | 0 | 0 | (0) | 0.13 | 0 |
| Ever Incarcerated | 0 | 0 | 0 | (0) | 0 | 0 |
| Experienced Partner Violence | 0 | 0 | 0 | (0) | 0 | 0 |
| Exchanged Sex for Money | 0 | 0 | 0 | (20) | 0.12 | 0 |
| Attempted Suicide | 0.75 | 0.64 | 0.72 | (100) | 0.78 | 0.72 |
| Hospitalized for Mental Health Problems | 0 | 0 | 0 | (0) | -0.01 | 0 |
| Sexually Abused | 0.25 | 0 | 0.03 | (62) | 0.10 | 0 |
| Had Sex with Someone 5+ Years Older Before Age 16 | -0.14 | 0 | 0 | (0) | -0.19 | 0 |
| Ever Been Robbed | 0.30 | 0.11 | 0.11 | (98) | 0.22 | 0.11 |
| Seen Serious Injury or Death | 0.04 | 0 | 0 | (16) | 0.05 | 0 |
| Family Member was Murdered | 0.18 | 0.39 | 0.28 | (100) | 0.64 | 0.26 |
| Used Drugs During Last Sexual Encounter | 0 | 0 | 0 | (0) | -0.16 | 0 |
| Ever Smoked | 0.14 | 0.22 | 0.20 | (100) | 0.37 | 0.19 |
|  |  |  |  |  |  |  |
| *R2* training | .711 | .685 | .703 | | .711 | .703 |
| MSE training | 10.103 | 10.785 | 10.487 | | 10.206 | 10.508 |
| *R2* test | .604 | .609 | .651 | | .636 | .652 |
| MSE test | 11.801 | 12.441 | 11.655 | | 12.162 | 11.651 |
| Lambda () | 0.090 | 0.118 | 0.113 | | 0.004 | 28.184 |

*Notes*. 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