

mi: Missing Data Imputation and Diagnostics – Opening Windows into the Black Box

Yu-Sung Su, Andrew Gelman, Jennifer Hill and Masanao
Yajima

Columbia University, Columbia University, New York University and University of
California at Los Angeles

July 8, 2009



Most MI packages are black boxes!

- Since Rubin' pioneering work (1987) multiple imputation methods of non-response in surveys, the general statistical theory and framework for managing missing information has been well developed.
- Numerous software packages have been developed; yet, to a certain degree, **they are black boxes**.
- These software users must trust the imputation procedure without much control over what goes into it and without much understanding of what comes out.



Most MI packages are black boxes!

- Since Rubin' pioneering work (1987) multiple imputation methods of non-response in surveys, the general statistical theory and framework for managing missing information has been well developed.
- Numerous software packages have been developed; yet, to a certain degree, **they are black boxes**.
- These software users must trust the imputation procedure without much control over what goes into it and without much understanding of what comes out.



Most MI packages are black boxes!

- Since Rubin' pioneering work (1987) multiple imputation methods of non-response in surveys, the general statistical theory and framework for managing missing information has been well developed.
- Numerous software packages have been developed; yet, to a certain degree, **they are black boxes**.
- These software users must trust the imputation procedure without much control over what goes into it and without much understanding of what comes out.



Modeling and model checking shall not be neglected in MI!

- Modeling checking and other diagnostics are an important part of any statistical procedure.
- These examination are particularly important because of the inherent tension of multiple imputation.
- The model used for the imputations is not in general the same as the model used for the analysis.
- Our mi is an open-ended, open source package, not only to solve imputation problems, but also to develop and implement new ideas in modeling and model checking.
- Our mi uses the chain equations algorithm.



Modeling and model checking shall not be neglected in MI!

- Modeling checking and other diagnostics are an important part of any statistical procedure.
- These examination are particularly important because of the inherent tension of multiple imputation.
- The model used for the imputations is not in general the same as the model used for the analysis.
- Our mi is an open-ended, open source package, not only to solve imputation problems, but also to develop and implement new ideas in modeling and model checking.
- Our mi uses the chain equations algorithm.



Modeling and model checking shall not be neglected in MI!

- Modeling checking and other diagnostics are an important part of any statistical procedure.
- These examination are particularly important because of the inherent tension of multiple imputation.
- The model used for the imputations is not in general the same as the model used for the analysis.
- Our mi is an open-ended, open source package, not only to solve imputation problems, but also to develop and implement new ideas in modeling and model checking.
- Our mi uses the chain equations algorithm.



Modeling and model checking shall not be neglected in MI!

- Modeling checking and other diagnostics are an important part of any statistical procedure.
- These examination are particularly important because of the inherent tension of multiple imputation.
- The model used for the imputations is not in general the same as the model used for the analysis.
- Our mi is an open-ended, open source package, not only to solve imputation problems, but also to develop and implement new ideas in modeling and model checking.
- Our mi uses the chain equations algorithm.



Novel Features in our MI package

- Bayesian regression models to address problems with separation;
- Imputation steps that deal with structural correlation;
- Functions that check the convergence of the imputations;
- Plotting functions that visually check the imputation models.



Novel Features in our MI package

- Bayesian regression models to address problems with separation;
- Imputation steps that deal with structural correlation;
- Functions that check the convergence of the imputations;
- Plotting functions that visually check the imputation models.



Novel Features in our MI package

- Bayesian regression models to address problems with separation;
- Imputation steps that deal with structural correlation;
- Functions that check the convergence of the imputations;
- Plotting functions that visually check the imputation models.



Novel Features in our MI package

- Bayesian regression models to address problems with separation;
- Imputation steps that deal with structural correlation;
- Functions that check the convergence of the imputations;
- Plotting functions that visually check the imputation models.



Outline

- 1 Introduction
- 2 Basic Setup of mi
- 3 Novel Features
- 4 Concluding Remarks



The Basic Procedure of Doing a Sensible MI

- Setup
 - Display of missing data patterns.
 - Identifying structural problems in the data and preprocessing.
 - Specifying the conditional models.
 - Iterative imputation based on the conditional model.
- Imputation
 - Checking the fit of conditional models.
 - Checking the convergence of the procedure.
 - Checking to see if the imputed values are reasonable.
- Analysis
 - Obtaining completed data.
 - Pooling the complete case analysis on multiply imputed datasets.
- Validation
 - Assessing the quality of imputation.
 - Cross-validation.



The Basic Procedure of Doing a Sensible MI

- Setup
 - Display of missing data patterns.
 - Identifying structural problems in the data and preprocessing.
 - Specifying the conditional models.
 - Iterative imputation based on the conditional model.
- Imputation
 - Checking the fit of conditional models.
 - Checking the convergence of the procedure.
 - Checking to see if the imputed values are reasonable.
- Analysis
 - Obtaining completed data.
 - Pooling the complete case analysis on multiply imputed datasets.
- Validation
 - Sensitivity analysis.
 - Cross validation.

The Basic Procedure of Doing a Sensible MI

- Setup
 - Display of missing data patterns.
 - Identifying structural problems in the data and preprocessing.
 - Specifying the conditional models.
 - Iterative imputation based on the conditional model.
- Imputation
 - Checking the fit of conditional models.
 - Checking the convergence of the procedure.
 - Checking to see if the imputed values are reasonable.
- Analysis
 - Obtaining completed data.
 - Pooling the complete case analysis on multiply imputed datasets.
- Validation
 - Sensitivity analysis.
 - Cross validation.



The Basic Procedure of Doing a Sensible MI

- Setup
 - Display of missing data patterns.
 - Identifying structural problems in the data and preprocessing.
 - Specifying the conditional models.
 - Iterative imputation based on the conditional model.
- Imputation
 - Checking the fit of conditional models.
 - Checking the convergence of the procedure.
 - Checking to see if the imputed values are reasonable.
- Analysis
 - Obtaining completed data.
 - Pooling the complete case analysis on multiply imputed datasets.
- Validation
 - Sensitivity analysis.
 - Cross validation.



Imputation Information Matrix

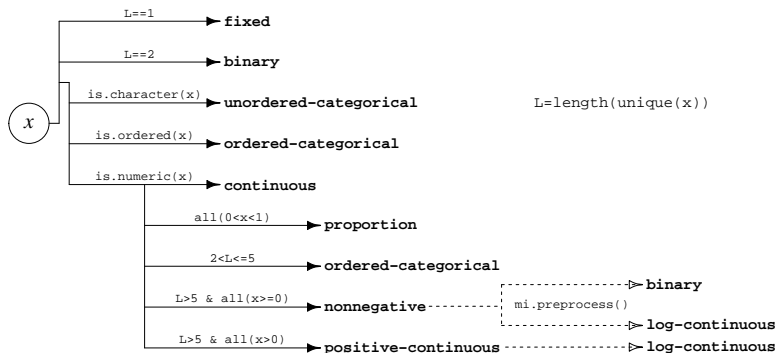
	names	include	order	number.mis	all.mis	type	collinear
1	h39b.W1	Yes	1	179	No	nonnegative	No
2	age.W1	Yes	2	24	No	positive-continuous	No
3	c28.W1	Yes	3	38	No	positive-continuous	No
4	pcs.W1	Yes	4	24	No	positive-continuous	No
5	mcs37.W1	Yes	5	24	No	binary	No
6	b05.W1	Yes	6	63	No	ordered-categorical	No
7	haartadhere.W1	Yes	7	24	No	ordered-categorical	No

```

                                h39b.W1
"haartadhere.W1 ~ age.W1 + c28.W1 + pcs.W1 + mcs37.W1 + b05.W1 + haartadhere.W1"
                                age.W1
"age.W1 ~ h39b.W1 + c28.W1 + pcs.W1 + mcs37.W1 + b05.W1 + haartadhere.W1"
                                c28.W1
"c28.W1 ~ h39b.W1 + age.W1 + pcs.W1 + mcs37.W1 + b05.W1 + haartadhere.W1"
                                pcs.W1
"pcs.W1 ~ h39b.W1 + age.W1 + c28.W1 + mcs37.W1 + b05.W1 + haartadhere.W1"
                                mcs37.W1
"mcs37.W1 ~ h39b.W1 + age.W1 + c28.W1 + pcs.W1 + b05.W1 + haartadhere.W1"
                                b05.W1
"b05.W1 ~ h39b.W1 + age.W1 + c28.W1 + pcs.W1 + mcs37.W1 + haartadhere.W1"
                                haartadhere.W1
"haartadhere.W1 ~ h39b.W1 + age.W1 + c28.W1 + pcs.W1 + mcs37.W1 + b05.W1"

```

Variable Types



Imputation Models

List of regression models, corresponding to variable types

Variable Types	Regression Models
binary	<code>mi.binary</code>
continuous	<code>mi.continuous</code>
count	<code>mi.count</code>
fixed	<code>mi.fixed</code>
log-continuous	<code>mi.continuous</code>
nonnegative	<code>mi.continuous</code>
ordered-categorical	<code>mi.polr</code>
unordered-categorical	<code>mi.categorical</code>
positive-continuous	<code>mi.continuous</code>
proportion	<code>mi.continuous</code>
predictive-mean-matching	<code>mi.pmm</code>



Bayesian Models to Address Problems with Separation

- The problem of separation occurs whenever the outcome variable is perfectly predicted by a predictor or a linear combination of several predictors.
- The problem exacerbates when the number of predictors increases.
- However, multiple imputation is generally strengthened by including many variables to help to satisfy the missing at random assumption.
- To address this problem of separation, we have augmented our mi to allow for Bayesian versions of generalized linear models (Gelman, Jakulin, Pittau and Su 2008).



Bayesian Models to Address Problems with Separation

List of Bayesian Generalized Linear Models, Used in Mi Regression Functions

mi Functions	Bayesian Functons
<code>mi.continuous()</code>	<code>bayesglm()</code> with gaussian family
<code>mi.binary()</code>	<code>bayesglm()</code> with binomial family (default uses logit link)
<code>mi.count()</code>	<code>bayesglm()</code> with quasi-poisson family (overdispersed poisson)
<code>mi.polr()</code>	<code>bayespolr()</code>



Imputing Semi-Continuous Data with Transformation

- positive-continuous, nonnegative and proportion variable types are, in general, not modeled in a reasonable way in other imputation software.
- These kinds of data have bounds or truncations and are not of standard distribution
- Our algorithm models these data with transformation.



Imputing Semi-Continuous Data with Transformation

- positive-continuous, nonnegative and proportion variable types are, in general, not modeled in a reasonable way in other imputation software.
- These kinds of data have bounds or truncations and are not of standard distribution
- Our algorithm models these data with transformation.
 - nonnegative: create two ancillary variables, one indicator for values bigger than 0; the other takes log of values bigger than 0.
 - positive-continuous: log of such a variable.
 - proportion: $\text{logit}(x)$



Imputing Semi-Continuous Data with Transformation

- positive-continuous, nonnegative and proportion variable types are, in general, not modeled in a reasonable way in other imputation software.
- These kinds of data have bounds or truncations and are not of standard distribution
- Our algorithm models these data with transformation.
 - nonnegative: create two ancillary variables, one indicator for values bigger than 0; the other takes log of values bigger than 0.
 - positive-continuous: log of such a variable.
 - proportion: $\text{logit}(x)$




Imputing Semi-Continuous Data with Transformation

- positive-continuous, nonnegative and proportion variable types are, in general, not modeled in a reasonable way in other imputation software.
- These kinds of data have bounds or truncations and are not of standard distribution
- Our algorithm models these data with transformation.
 - nonnegative: create two ancillary variables, one indicator for values bigger than 0; the other takes log of values bigger than 0.
 - positive-continuous: log of such a variable.
 - proportion: $\text{logit}(x)$



Imputing Semi-Continuous Data with Transformation

Original Dataset				Transformed Dataset			
x1	x2	x3		x1	x1.ind	x2	x3
0.00	6.00	NA		NA	0.00	1.79	NA
NA	5.00	0.18		NA	NA	1.61	-1.49
6.00	10.00	0.54		1.79	1.00	2.30	0.15
19.00	NA	0.51	<code>mi.preprocess()</code>	2.94	1.00	NA	0.04
0.00	NA	0.43		NA	0.00	NA	-0.27
5.00	NA	0.98		1.61	1.00	NA	4.00
NA	11.00	0.26		NA	NA	2.40	-1.06
10.00	NA	0.82		2.30	1.00	NA	1.54
18.00	5.00	0.16		2.89	1.00	1.61	-1.65
0.00	14.00	NA		NA	0.00	2.64	NA
...

Imputing Data with Collinearity

- Perfect Correlation: $x_1 = 10x_2 - 5$
 - Same missing pattern, mi duplicates values from the correlated variable.
 - Different missing pattern, mi adds noises to the iterative imputation process (see next point).
- Additive constraints: $x_1 = x_2 + x_3 + x_4 + x_5 + 10$
 - Reshuffling noise: mi randomly imputes missing data from the marginal distribution, not from the conditional models
 - Fading empirical noise: mi augments the data (drawn from the observed data) by 10% of the completed data.



Imputing Data with Collinearity

- Perfect Correlation: $x_1 = 10x_2 - 5$
 - Same missing pattern, mi duplicates values from the correlated variable.
 - Different missing pattern, mi adds noises to the iterative imputation process (see next point).
- Additive constraints: $x_1 = x_2 + x_3 + x_4 + x_5 + 10$
 - Reshuffling noise: mi randomly imputes missing data from the marginal distribution, not from the conditional models
 - Fading empirical noise: mi augments the data (drawn from the observed data) by 10% of the completed data.
 - After one batch of finished mi, mi will run 20 more iteration without noises to alleviate the influence of the noise.



Imputing Data with Collinearity

- Perfect Correlation: $x_1 = 10x_2 - 5$
 - Same missing pattern, mi duplicates values from the correlated variable.
 - Different missing pattern, mi adds noises to the iterative imputation process (see next point).
- Additive constraints: $x_1 = x_2 + x_3 + x_4 + x_5 + 10$
 - Reshuffling noise: mi randomly imputes missing data from the marginal distribution, not from the conditional models
 - Fading empirical noise: mi augments the data (drawn from the observed data) by 10% of the completed data.
 - After one batch of finished mi, mi will run 20 more iteration without noises to alleviate the influence of the noise.



Imputing Data with Collinearity

- Perfect Correlation: $x_1 = 10x_2 - 5$
 - Same missing pattern, mi duplicates values from the correlated variable.
 - Different missing pattern, mi adds noises to the iterative imputation process (see next point).
- Additive constraints: $x_1 = x_2 + x_3 + x_4 + x_5 + 10$
 - Reshuffling noise: mi randomly imputes missing data from the marginal distribution, not from the conditional models
 - Fading empirical noise: mi augments the data (drawn from the observed data) by 10% of the completed data.
 - After one batch of finished mi, mi will run 20 more iteration without noises to alleviate the influence of the noise.



Imputing Data with Collinearity

- Perfect Correlation: $x_1 = 10x_2 - 5$
 - Same missing pattern, mi duplicates values from the correlated variable.
 - Different missing pattern, mi adds noises to the iterative imputation process (see next point).
- Additive constraints: $x_1 = x_2 + x_3 + x_4 + x_5 + 10$
 - Reshuffling noise: mi randomly imputes missing data from the marginal distribution, not from the conditional models
 - Fading empirical noise: mi augments the data (drawn from the observed data) by 10% of the completed data.
 - After one batch of finished mi, mi will run 20 more iteration without noises to alleviate the influence of the noise.



Imputing Data with Collinearity

Structured Correlated Dataset

class	total	female	male
1	66	NA	35
2	NA	27	23
3	76	37	NA
4	NA	31	NA
5	51	24	NA
6	73	39	34
7	NA	NA	39
8	NA	NA	NA
9	46	26	NA
10	59	NA	NA
...



Checking the Convergence of the Imputation

- mi monitors the mixing of each variable by the variance of its mean and standard deviation within and between different chains of the imputation ($\hat{R} < 1.1$ (Gelman et al 2004))
- By default, mi monitors the convergence of the mean and standard deviation of the imputed data.
- More strictly, mi can also monitor the convergence of the parameters of the conditional models.



Checking the Convergence of the Imputation

- mi monitors the mixing of each variable by the variance of its mean and standard deviation within and between different chains of the imputation ($\hat{R} < 1.1$ (Gelman et al 2004))
- By default, mi monitors the convergence of the mean and standard deviation of the imputed data.
- More strictly, mi can also monitor the convergence of the parameters of the conditional models.



Checking the Convergence of the Imputation

- mi monitors the mixing of each variable by the variance of its mean and standard deviation within and between different chains of the imputation ($\hat{R} < 1.1$ (Gelman et al 2004))
- By default, mi monitors the convergence of the mean and standard deviation of the imputed data.
- More strictly, mi can also monitor the convergence of the parameters of the conditional models.



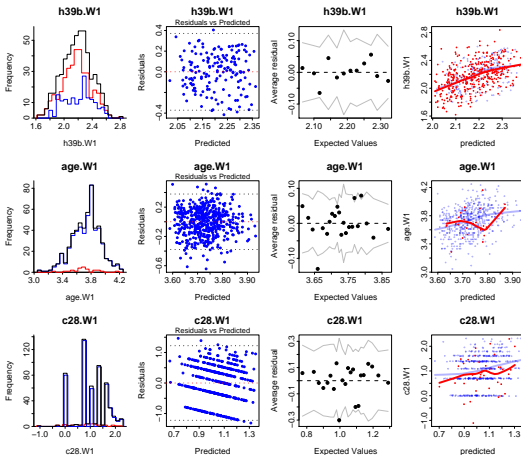
Model checking and Other Diagnostic for the Imputation Using Graphics

We offers three strategies:

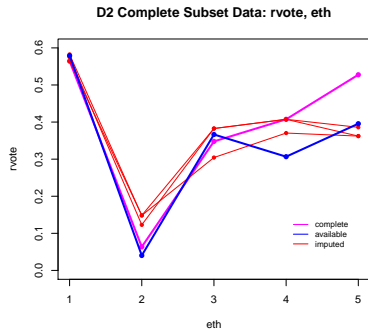
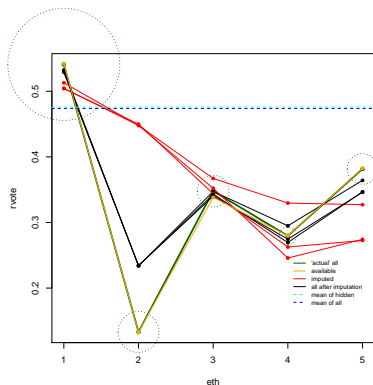
- Imputations are typically generated using models, such as regressions or multivariate distributions, which are fit to observed data. Thus the fit of these models can be checked (Gelman 2005).
- Imputations can be checked using a standard of reasonability: the differences between observed and missing values, and the distribution of the completed data as a whole, can be checked to see whether they make sense in the context of the problem being studied (Abayomi et al 2008).
- We can use cross-validation to perform sensitivity analysis to violations of our assumptions.



Model checking and Other Diagnostic for the Imputation Using Graphics



Model checking and Other Diagnostic for the Imputation Using Graphics



The Contributions of our MI package

- 1** Graphical diagnostics of imputation models and convergence of the imputation process.
- 2 Use of Bayesian version of regression models to handle the issue of separation.
- 3 Imputation model specification is similar to the way in which you would fit a regression model in R
- 4 Automatical detection of problematic characteristics fo data followed by either a fix or an alert to the user. In particular, mi add noise into the imputation process to solve the problem of additive constraints.



The Contributions of our MI package

- 1** Graphical diagnostics of imputation models and convergence of the imputation process.
- 2** Use of Bayesian version of regression models to handle the issue of separation.
- 3** Imputation model specification is similar to the way in which you would fit a regression model in R
- 4** Automatical detection of problematic characteristics fo data followed by either a fix or an alert to the user. In particular, mi add noise into the imputation process to solve the problem of additive constraints.



The Contributions of our MI package

- 1** Graphical diagnostics of imputation models and convergence of the imputation process.
- 2** Use of Bayesian version of regression models to handle the issue of separation.
- 3** Imputation model specification is similar to the way in which you would fit a regression model in R
- 4** Automatical detection of problematic characteristics fo data followed by either a fix or an alert to the user. In particular, mi add noise into the imputation process to solve the problem of additive constraints.



The Contributions of our MI package

- 1 Graphical diagnostics of imputation models and convergence of the imputation process.
- 2 Use of Bayesian version of regression models to handle the issue of separation.
- 3 Imputation model specification is similar to the way in which you would fit a regression model in R
- 4 Automatical detection of problematic characteristics fo data followed by either a fix or an alert to the user. In particular, mi add noise into the imputation process to solve the problem of additive constraints.



The Future Plan

- Speed up!
- Modeling time-series cross-sectional data, hierarchical or clustered data.



Link to the Full Paper

Yu-Sung Su, Andrew Gelman, Jennifer Hill, and Masanao Yajima.
“Multiple imputation with diagnostics (mi) in R: Opening windows into the black box”. *Journal of Statistical Software*. Available at:
<http://www.stat.columbia.edu/~gelman/research/published/mipaper.rev04.pdf>

