Multiple Imputation of Survey Data

Overview and Application

Kasi Laxmanan

Table of contents

Missing Data - What(Where) is it?

- Missing data - Data that are missing for some (but not all) variables and for some (but not all) cases/observations.
- Latent or Unobserved: Data are missing on a variable for all cases.
- Unit Non-response: Data are missing on all variables for some cases.
- Item Non-response: a.k.a Missing Data - Topic of discussion.
- Goal: Most work on missing data presumes that the goal is to get optimal estimates of parameters (e.g., unbiased and efficient).

Table of contents

Missing Data - Some Definitions

Missing completely at random (MCAR)

- The probability that a particular variable is missing for a particular individual does not depend on the value of any variables in the model of interest.
- There is no relationship between the missingness of the data and any values, observed or missing. Those missing data points are a random subset of the data. There is nothing systematic going on that makes some data more likely to be missing than others.

Example:

Consider a data set with administrative records and personal interviews, if the subject declines to be interviewed, all the responses in that interview will be jointly missing.

It typically occurs when data sets are pieced together from multiple sources.

Missing Data - Some Definitions

Missing at Random (MAR)

- Whether an observation is missing has nothing to do with the missing values, but it does have to do with the values of an individual's observed variables.
- MCAR vs MAR- [1] Little's test. [2] To create dummy variables for whether a variable is missing(1 = missing; 0 = observed), then run t-tests and chi-square tests between this variable and other variables.
- For example, if women really are less likely to tell you their weight than men, a chi-square test will tell you that the percentage of missing data on the weight variable is higher for women than men.
- "Missing Completely at Random" and "Missing at Random" are both considered 'ignorable'.

Example:

If men are more likely to tell you their weight than women, weight is MAR.

Missing Data - Some Definitions

Not missing at random (NMAR) a.k.a MNAR(Missing-not-at-Random)

- The probability that a variable is missing depends on the (unknown) value of that variable, after adjusting for other variables in the model.
- There is a relationship between the propensity of a value to be missing and its values.
- NMAR is called “non-ignorable” because the missing data mechanism itself has to be modeled as you deal with the missing data. You have to include some model for why the data are missing and what the likely values are.
- NMAR vs MAR: follow-up on a paper survey with phone calls to a group of the non-respondents and ask a few key survey items. This allows you to compare respondents to non-respondents.If key items differ by very much, they are NMAR.

Example:

The people with the lowest education are missing on education or the sickest people are most likely to drop out of the study.

Table of contents

Methods for Handling Missing Data

Criteria for evaluating Missing-data methods

1. Minimize bias. Although it is well-known that missing data can introduce bias into parameter estimates, a good method should make that bias as small as possible.
2. Maximize the use of available information. We want to avoid discarding any data, and we want to use the available data to produce parameter estimates that are efficient (i.e.,have minimum sampling variability).
3. Yield good estimates of uncertainty. We want accurate estimates of standard errors, confidence intervals and p-values.
4. In addition, it would be nice if the missing-data method could accomplish these goals without making unnecessarily restrictive assumptions about the missing-data mechanism.

Conventional Methods

List-wise Deletion:

- Delete cases with any missing data on the variables of interest.
- Requires the data are MCAR in order to not introduce bias in the results.
- Pros: Is good on 3 (above), terrible on 2 and so-so on 1. For MCAR, it does not introduce bias. If the data are MAR but not MCAR, list-wise deletion may introduce bias.
- Cons: Data Loss; Larger standard errors, Wider confidence intervals, and a loss of power in testing hypotheses.

Pair-wise Deletion:

- Each of the 'moments' is estimated for each variable or each pair of variables; sample moments are substituted for the population parameters.
- Pros: All data is used. Cons: If the data are MAR but not MCAR, may yield biased estimates. Does not produce accurate estimates.

Conventional Methods -Contd

Dummy-variable adjustment:

- A dummy variable is created to indicate whether or not data are missing on that predictor. All such dummy variables are included as predictors in the regression. Cases with missing data on a predictor are coded, usually with the mean for non- missing cases.
- Pros: All available data is used.
- Cons: Biased estimates even if MCAR.

Missing Imputation:

- A simple but popular approach is to substitute means(or median) for missing values.
- Cons: This is well-known to produce biased estimates, produce underestimates of standard errors, which in turn leads to inflated test statistics and p-values that are too low.

Survey Data: NLSY

- We will use a data set that has records for 581 children who were interviewed in 1990 as part of the National Longitudinal Survey of Youth (NLSY). The Original data did not have missing values, MCAR simulated dataset(Table 1) was used for the analysis.
- Goal: Estimate a linear-regression model with ANTI as the dependent variable and all the others as predictors.
- List-wise deletion leaves only 225 cases.

Table 1

Variable	Label	N	N Miss	%Missing	Minimum	Mean	Maximum
anti	child antisocial behavior scale 0-6	581	0	0%	0	1.5679862	6
self	child self-esteem scale 6-24	433	148	25%	9	20.0508083	24
pov	family poverty status	431	150	26%	0	0.3480278	1
black	child race black	468	113	19%	0	0.3589744	1
hispanic	child race hispanic	468	113	19%	0	0.2435897	1
childage	child age	581	0	0%	8	8.9436317	10
divorce	mother divorced	581	0	0%	0	0.2358003	1
gender	childs gender: 1-female 0-male	581	0	0%	0	0.5043029	1
momage	mother age at birth of child	581	0	0%	16	20.6557659	25
momwork	mother employment status	495	86	15%	0	0.3353535	1
race	child race	468	113	19%	1	1.8461538	3
povq	family poverty status	431	150	26%	0	0.3480278	1

Maximum Likelihood

- The best available alternatives are maximum likelihood and multiple imputation.
- If the assumptions are met (MAR and any distributional assumptions), both produce estimates that are consistent, asymptotically efficient (or close to it), and asymptotically normal.
- Under MAR, maximum likelihood is implemented as follows: for each individual with missing data, integrate the likelihood function over the joint distribution of variables with missing data.
- Multiply the likelihoods for all individuals to construct the overall likelihood.
- For missing data on predictors, PROC CALIS can do ML for a large class of linear models. You need to use the option METHOD=FIML on the PROC statement. FIML stands for full information maximum likelihood. Assumes multivariate normality for variables with missing data.
- For logistic regression, Mplus is the only commercial software package that will handle missing data by maximum likelihood(as far as my research goes).
- For this study, we did not focus on an applied solution for Maximum Likelihood.

Multiple Imputation

- Will focus on multiple imputation using PROC MI in SAS, which assumes that data are missing at random.
- In multiple imputation, the imputed values are random draws from the conditional distribution of the variables with missing data, given the observed data.
- Multiple data sets are produced (25 is the default in PROC MI—used to be 5)
- The model of interest is estimated on each data set, producing approximately unbiased estimates and the standard errors are low.
- The results are combined using some simple rules (Rubin 1987). 1) Just take the unweighted mean of the parameter estimates. 2) Standard errors combine variance within data sets and variance between data sets.

Figure 1: SAS Code for MI

```
proc mi data=prob1 out=miout
nimpute=15;
var anti self pov black hispanic
divorce gender momwork;
proc reg data=miout outest=a covout;
model anti=self pov black hispanic
divorce gender momwork;
by _imputation_;
proc mianalyze data=a;
var intercept self pov black
hispanic divorce gender momwork;
run;
```

Multiple Imputation - Output

- Every variable with missing data is imputed based on regressions with all the other variables as predictors, along with random draws from the error terms.
- Although we don't impute the dependent variable, it's essential that the dependent variable be included so that it serves as a basis to impute the predictors.
- PROC MI produces 25 data sets that are stacked into one data set, MIOUT.
- Standard errors for most coefficients are considerably smaller than with list-wise deletion. DF is high for all coefficients, a good sign. If DF <100, that is an indication that more imputed data sets are needed.
- Unless you set the seed for the random number generator (SEED= option on the PROC MI statement), results will be a little different every time you run the program.

Table 2: Outputs for MI

Parameter Estimates (15 Imputations)										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
intercept	3.021437	0.488453	2.05743	3.98544	175.32	2.578489	3.449793	0	6.19	<.0001
self	-0.068396	0.022015	-0.11182	-0.02497	193.15	-0.085323	-0.048330	0	-3.11	0.0022
pov	0.642829	0.176339	0.29258	0.99308	91.451	0.503677	0.844906	0	3.65	0.0004
black	0.079106	0.165588	-0.24744	0.40565	197.66	-0.063747	0.203445	0	0.48	0.6334
hispanic	-0.345425	0.171108	-0.68211	-0.00874	308.96	-0.504739	-0.209890	0	-2.02	0.0444
divorce	-0.111145	0.152412	-0.41027	0.18798	902.35	-0.192722	-0.033044	0	-0.73	0.4660
gender	-0.566282	0.117747	-0.79708	-0.33548	13975	-0.608028	-0.531811	0	-4.81	<.0001
momwork	0.210769	0.138021	-0.06020	0.48174	719.85	0.137705	0.289989	0	1.53	0.1272

Comparison of Methods

- 1st set of estimates is based on the original data set with no missing data. Three variables have p-values below .01: SELF, POV and GENDER. Higher levels of antisocial behavior are associated with lower levels of self-esteem, being in poverty and being male. The negative coefficient for Hispanic is also marginally significant.
- 2nd set of estimates was obtained with list-wise deletion. Although the coefficients are reasonably close to those in the original data set, the standard errors are much larger, reflecting the fact that more than half the cases are lost. As a result, only the coefficient for POV is statistically significant.
- ML and MI estimates are shown in the 3rd/4th panel. The coefficients are closer to the original values in list-wise deletion. More importantly, the estimated standard errors (using 431 as the sample size in the two- step method) are much lower than those from list-wise deletion, with the result that POV, SELF and GENDER all have p-values below .01.

Table 3

	No missing data		*Listwise deletion*		*Maximum likelihood*			*Multiple imputation*	
Variable	*Coeff.*	*SE*	*Coeff.*	*SE*	*Coeff.*	*Two-step SE*	*Direct SE*	*Coeff.*	*SE*
SELF	**−0.054**	0.018	−0.045	0.031	**−0.066**	0.022	0.022	**−0.069**	0.021
POV	**0.565**	0.137	**0.727**	0.234	**0.635**	0.161	0.162	**0.625**	0.168
BLACK	0.090	0.140	0.053	0.247	0.071	0.164	0.160	0.073	0.155
HISPANIC	−0.346	0.153	−0.353	0.253	−0.336	0.176	0.170	−0.332	0.168
DIVORCE	0.068	0.144	0.085	0.243	−0.109	0.166	0.146	−0.107	0.147
GENDER	**−0.537**	0.117	−0.334	0.197	**−0.560**	0.135	0.117	**−0.556**	0.118
MOMWORK	0.184	0.129	0.259	0.216	0.215	0.150	0.142	0.242	0.143

Coefficients (Coeff.) in bold are statistically significant at the .01 level.
SE, standard error.

Misconceptions on Missing Data

True/False

1. Imputation is really just making up data to artificially inflate results. It's better to just drop cases with missing data than to impute.

Misconceptions on Missing Data

True/False

1. Imputation is really just making up data to artificially inflate results. It's better to just drop cases with missing data than to impute.
   Answer: False!
   Multiple imputation, when done well, gives pretty much the same unbiased results, with full power, as the full non-missing data set.
2. We can just impute the mean for any missing data. It won't affect results, and improves power.

Misconceptions on Missing Data

True/False

1. Imputation is really just making up data to artificially inflate results. It's better to just drop cases with missing data than to impute.
   Answer: False!
   Multiple imputation, when done well, gives pretty much the same unbiased results, with full power, as the full non-missing data set.
2. We can just impute the mean for any missing data. It won't affect results, and improves power.
   Answer; False!
   Mean imputation is bad imputation. It does improve power, but your results will be so biased, the improved power won't help much.
3. Missing data isn't really a problem if we just doing simple statistics, like chi-squares and t-tests.

Misconceptions on Missing Data

True/False

1. Imputation is really just making up data to artificially inflate results. It's better to just drop cases with missing data than to impute.
   Answer: False!
   Multiple imputation, when done well, gives pretty much the same unbiased results, with full power, as the full non-missing data set.
2. We can just impute the mean for any missing data. It won't affect results, and improves power.
   Answer; False!
   Mean imputation is bad imputation. It does improve power, but your results will be so biased, the improved power won't help much.
3. Missing data isn't really a problem if we just doing simple statistics, like chi-squares and t-tests.
   Answer; False!
   The percent, pattern, and randomness of the missing data that determines how problematic missing data are. Even simple statistics need to be accurate and unbiased.
4. The worst thing that missing data does is lower sample size and reduce power.

Misconceptions on Missing Data

True/False

1. Imputation is really just making up data to artificially inflate results. It's better to just drop cases with missing data than to impute.
   Answer: False!
   Multiple imputation, when done well, gives pretty much the same unbiased results, with full power, as the full non-missing data set.
2. We can just impute the mean for any missing data. It won't affect results, and improves power.
   Answer; False!
   Mean imputation is bad imputation. It does improve power, but your results will be so biased, the improved power won't help much.
3. Missing data isn't really a problem if we just doing simple statistics, like chi-squares and t-tests.
   Answer; False!
   The percent, pattern, and randomness of the missing data that determines how problematic missing data are. Even simple statistics need to be accurate and unbiased.
4. The worst thing that missing data does is lower sample size and reduce power.
   Answer; False!
   List wise deletion is bad. But even worse are biased parameter estimates and biased standard errors.

Emerging Techniques -Multiple Imputation Chained Equations(MICE)

- PROC MI has an alternative imputation method called the fully conditional specification (FCS), also known as MICE (multiple imputation for chained equations).
- In FCS, you can specify a different imputation model for each variable with missing data.
- These models can be based on linear regression, logistic regression (either binary or multinomial), discriminant analysis, or predictive mean matching.
- In R, using mice package we can specify the distributions for variables with method variable and exclude variable(s) in predictorMatrix, run the mice function with maxit(no of iterations), then run lm function and summary to look at estimates. In SAS a similar approach was used with FCS statement.

Figure 2: mice package in R vs SAS PROC MI

```r
# We run the mice code with 0 iterations
imp <- mice(nlsy, maxit=0)
imp
# Extract predictorMatrix and methods of imputation
predM = imp$predictorMatrix
nmeth = imp$method

# With this command, we tell mice to impute the anesimp2 data, create 5
# datasets, use predM as the predictor matrix and don't print the imputation
# process. If you would like to see the process, set print as TRUE

imp2 <- mice(nlsy, maxit = 5,
             predictorMatrix = predM,
             method = nmeth, print =  FALSE)

# First, turn the datasets into long format
nlsy_long <- mice::complete(imp2, action="long", include = TRUE)

# Convert back to mids type - mice can work with this type
nlsy_long_mids<-as.mids(nlsy_long)
# Regression
fitimp <- with(nlsy_long,
               lm(anti2 ~ self2+pov2+black+hispanic+gender+momwork))
summary((fitimp))
```

```
PROC MI DATA=prob1 OUT=miout2;
var anti self pov black hispanic
divorce gender momwork;
FCS REG(self pov black hispanic
divorce gender momwork);
RUN;

proc reg data=miout2 outest=mice covout;
model anti=self pov black hispanic
divorce gender momwork;
by _imputation_;
run;

proc mianalyze data=mice;
var intercept self pov black
hispanic divorce gender momwork;
run;
```

Multiple Imputation Chained Equations -Contd

- In the NLSY data, we find that overall 38 percent non-missing values; black and hispanic contributed to 8 percent missing; 13 percent for poverty status and 12 percent for self esteem. See Table 4.
- Next, we used FCS statement to specify the distribution of variables, we chose the default number of iterations which ran 25 imputations, then we ran Regression with all 25 iterations and used PROC MIANALYZE to look at the estimates.
- As seen in Multiple Imputation, the coefficients were closer to the original values in list-wise deletion, the estimated standard errors were also much lower than those from list-wise deletion, with the result that POV, SELF and GENDER still had p-values below .01.

Table 4: Outputs from MICE

Missing Data Patterns

Group	anti	self	pov	black	hispanic	divorce	gender	momwork	Freq	Percent	Group Means: anti	self	pov	black	hispanic	divorce	gender	momwork
1	X	X	X	X	X	X	X	X	225	38.73	1.880000	19.893333	0.382222	0.377778	0.266667	0.226667	0.622222	0.364444
2	X	X	X	X	X	X	X	.	41	7.06	1.951220	20.390244	0.439024	0.341463	0.121951	0.243902	0.536585	.
3	X	X	X	.	.	X	X	X	48	8.26	1.895833	20.625000	0.437500	.	.	0.291667	0.541667	0.312500
4	X	X	X	.	.	X	X	.	8	1.38	2.000000	18.500000	0.375000	.	.	0.250000	0.375000	.
5	X	X	.	X	X	X	X	X	77	13.25	1.467532	20.090909	.	0.480519	0.194805	0.259740	0.545455	0.324675
6	X	X	.	X	X	X	X	.	8	1.38	2.375000	20.000000	.	0.375000	0.125000	0.250000	0.375000	.
7	X	X	.	.	.	X	X	X	22	3.79	1.272727	20.272727	.	.	.	0.227273	0.636364	0.272727
8	X	X	.	.	.	X	X	.	4	0.69	1.000000	19.750000	.	.	.	0.250000	0.500000	.
9	X	.	X	X	X	X	X	X	71	12.22	1.197183	.	0.225352	0.267606	0.295775	0.239437	0.323944	0.323944
10	X	.	X	X	X	X	X	.	13	2.24	0.992308	.	0.076923	0.230769	0.230769	0.153846	0.692308	.
11	X	.	X	.	.	X	X	X	22	3.79	0.863636	.	0.227273	.	.	0.227273	0.500000	0.363636
12	X	.	X	.	.	X	X	.	3	0.52	0.666667	.	0	.	.	0.333333	0.333333	.
13	X	.	.	X	X	X	X	X	24	4.13	1.375000	20.083333	.	0.208333	0.291667	0.125000	0.583333	0.291667
14	X	.	.	X	X	X	X	.	9	1.55	0.888889	.	.	0.222222	0.222222	0.444444	0.444444	.
15	X	.	.	.	.	X	X	X	6	1.03	0.833333	.	.	.	.	0	1.000000	0

Parameter Estimates (25 Imputations)

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
intercept	2.980519	0.486118	2.02393	3.93711	303.12	2.501226	3.590782	0	6.13	<.0001
self	-0.067066	0.022223	-0.11080	-0.02333	291.31	-0.095543	-0.045350	0	-3.02	0.0028
pov	0.606365	0.178698	0.25318	0.95955	145.29	0.392401	0.800837	0	3.39	0.0009
black	0.106729	0.162439	-0.21267	0.42613	376.93	-0.046922	0.267002	0	0.66	0.5116
hispanic	-0.316871	0.167688	-0.64611	0.01237	687.43	-0.458450	-0.195425	0	-1.89	0.0592
divorce	-0.104099	0.148866	-0.39595	0.18775	4342.8	-0.178743	-0.012692	0	-0.70	0.4844
gender	-0.565258	0.117622	-0.79580	-0.33471	30274	-0.617235	-0.521890	0	-4.81	<.0001
momwork	0.230350	0.149327	-0.06335	0.52405	348.36	0.113338	0.387602	0	1.54	0.1238

Unsupervised Machine Learning for Imputation

- K-nearest neighbors (KNN) is an unsupervised locality-based regression and classification method.
- It considers each row of data as a point in n-dimensional space and finds k similar (neighboring) points based on their distance (for example, Euclidean for numeric data and Hamming for categorical data).
- To find the value for that row and that column of data, it takes an average of all its neighboring rows for that column and assigns the average as a value.
- For NLSY data, Python was used to load the data, impute using KNNImputer and run a linear regression using StatsModels modules using the GOOGLE COLAB environment.

Figure 3

```
[ ] nlsy=pd.read_sas('/content/nlsymiss.sas7bdat')

[ ] from sklearn.impute import KNNImputer
    imputer = KNNImputer()
    nlsy_i = imputer.fit_transform(nlsy)

[ ] nlsy
```

Figure 4

```
[ ] nlsy_imp.describe()
```

	anti	self	pov	black	hispanic	childage	divorce	gender	momage	momwork	race	povq
count	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000	581.000000
mean	1.567986	20.207401	0.329604	0.339931	0.252151	8.943632	0.235800	0.504303	20.655766	0.328744	1.844234	0.329604
std	1.470728	2.879253	0.444566	0.454126	0.412749	0.601355	0.424864	0.500412	2.188982	0.455257	0.758462	0.444566
min	0.000000	9.000000	0.000000	0.000000	0.000000	8.000000	0.000000	0.000000	16.000000	0.000000	1.000000	0.000000
25%	0.000000	19.000000	0.000000	0.000000	0.000000	8.416667	0.000000	0.000000	19.000000	0.000000	1.000000	0.000000
50%	1.000000	21.000000	0.000000	0.000000	0.000000	8.916667	0.000000	1.000000	21.000000	0.000000	2.000000	0.000000
75%	2.000000	22.000000	1.000000	1.000000	0.500000	9.416667	0.000000	1.000000	22.000000	1.000000	2.000000	1.000000

K-nearest neighbors (KNN) for Imputation

- KNN was used for imputing missing values. The intuition is that missing value should have a value similar to its neighboring points.
- First, the sas dataset was read into a pandas dataframe. The KNNImputer from sklearn.impute module was imported and run imputing missing values. See Figure 3.
- Next, the respective list of columns were loaded into vectors. Then, statsmodels.api module was used to fit a linear regression model and regression estimates are presented in the Figure 5 below.
- The standard error of the estimates in comparison to MI are same or smaller for self, black, hispanic, divorce, gender and momwork. Variables - self, pov, and gender all have p-values less than 0.01, comparable to MI.
- The coefficients are slightly different than what is reported from MI.

Figure 5: KNN impute and OLS using Statsmodels

```
[24]  1  nlsy_imp=pd.DataFrame(data=nlsy_i[0:,0:],
      2                 index=[i for i in range(nlsy_i.shape[0])],
      3                 columns=['anti','self','pov','black','hispanic','childage',
      4                          'divorce','gender','momage','momwork','race','povq'])

[39]  1
      2  x=nlsy_imp[['self','pov','black','hispanic','divorce','gender','momwork']]
      3  y=nlsy_imp['anti']
      4

      1  import statsmodels.api as sm
      2  results = sm.OLS(y,x).fit()
      3  results.summary()
```

	coef	std err	t	P>|t|	[0.025	0.975]
self	0.0637	0.006	10.904	0	0.052	0.075
pov	0.8207	0.157	5.238	0	0.513	1.128
black	0.3252	0.153	2.129	0.034	0.025	0.625
hispanic	-0.027	0.164	-0.164	0.869	-0.349	0.295
divorce	0.0098	0.146	0.067	0.947	-0.276	0.296
gender	-0.4524	0.119	-3.805	0	-0.686	-0.219
momwork	0.2788	0.14	1.995	0.047	0.004	0.553

Summary

1. We reviewed some of the definitions associated with Missing Data Imputation.
2. We looked at some of the conventional methods for handling missing data: List-wise Deletion, Maximum Likelihood and Multiple Imputation etc.
3. We looked at comparison of the above methods.
4. We looked at some of the common misconceptions associated with Missing Imputation techniques.
5. Finally, we looked at some of the emerging techniques such as Multiple Imputation Chained Equations and KNN for Imputation.

References I


Allison, P.D.
Missing Data..
Thousand Oaks, CA: Sage, 2001.
Allison, P.D. (2001) Missing Data. Thousand Oaks, CA: Sage.


Ashish Kumar.
Mastering pandas - Second Edition - Section 3.
Packt Publishing, ISBN: 9781789343236, 2019.


Stef Van Burren.
MICE: Multivariate Imputation by Chained Equations in R
Journal of Statistical Software, MMMMMM YYYY, Volume VV, Issue II., 2015.

References II

- Karen Grace Martin.
  How to Diagnose the Missing Data Mechanism, 2020
  https://www.theanalysisfactor.com/missing-data-mechanism/
  https://www.theanalysisfactor.com/answers-to-the-missing-data-quiz/
- University of Virginia, Library et all.
  Getting Started with Multiple Imputation in R, 2019
  https://uvastatlab.github.io/2019/05/01/getting-started-with-multiple-imputation-in-r/
- Edwin Ponraj Thangarajan et all.
  Missing Data Imputation and Its Effect on the Accuracy of Classification, 2018
  https://www.lexjansen.com/phuse-us/2018/dh/DH04.pdf

References III

- Yang Liu et all.
  Multiple Imputation by Fully Conditional Specification for Dealing with Missing Data in a Large Epidemiologic Study, 2015
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4945131/#R10
- SAS Institute.
  SAS online Documentation for Proc MI, 2020
  https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_mi_sect008.htm
- Mike Molter, Wright Avenue.
  Python-izing the SAS Programmer, 2019
  https://www.pharmasug.org/proceedings/2019/AP/PharmaSUG-2019-AP-212.pdf