Paper 2081-2015

Multiple Imputation Using the Fully Conditional Specification Method: A Comparison of SAS®, Stata, IVEware, and R

Patricia A. Berglund, University of Michigan-Institute for Social Research

ABSTRACT

This presentation emphasizes use of SAS 9.4 to perform multiple imputation of missing data using the PROC MI Fully Conditional Specification (FCS) method with subsequent analysis using PROC SURVEYLOGISTIC and PROC MIANALYZE. The data set used is based on a complex sample design. Therefore, the examples correctly incorporate the complex sample features and weights. The demonstration is then repeated in Stata, IVEware, and R for a comparison of major software applications that are capable of multiple imputation using FCS or equivalent methods and subsequent analysis of imputed data sets based on a complex sample design.

INTRODUCTION

Paper 2081-2015 presents a detailed example of multiple imputation of missing data from a complex sample design with the Fully Conditional Specification (FCS) method of PROC MI with subsequent analysis using PROC SURVEYLOGISTIC and PROC MIANALYZE. The application is then repeated using Stata, IVEware, and R with an equivalent imputation method while also accounting for the complex sample design features. The application replications enable a direct comparison of results from each software tool.

Analysts will gain knowledge and practical guidance for correctly implementing the three step multiple imputation process using data derived from a complex sample design data set. A general overview of the multiple imputation process is included but it is assumed that the analyst has a basic understanding of the MI process and analysis of complex sample design data.

MULTIPLE IMPUTATION OF MISSING DATA

Multiple Imputation is a robust and flexible option for handling missing data. MI is implemented following a framework for estimation and inference based upon a three step process: 1) formulation of the imputation model and imputation of missing data using PROC MI with a selected method, 2) analysis of complete data sets using standard SAS procedures (that assume the data are identically and independently distributed or from a simple random sample) or SURVEY procedures for analysis of data from a complex sample design, and 3) analysis of the output from the two previous steps using PROC MIANALYZE (Berglund and Heeringa, 2014). A key assumption made in the MI and MIANALYZE procedures is that the missing data are missing at random (MAR) or in other words, the probability that an observation is missing depends on observed Y but not missing Y, (Rubin, 1987).

The featured application implements the PROC MI FCS method for imputation of missing data in Step 1, use of PROC SURVEYLOGISTIC to perform design-based logistic regression in Step 2, and PROC MIANALYZE to combine results from previous steps in Step 3.

THE FULLY CONDITIONAL SPECIFICATION (FCS) METHOD

The Fully Conditional Specification (FCS) method is widely used for imputation of missing data for large mixed sets of continuous, nominal, ordinal, count and semi-continuous variables. The FCS method is also labeled the sequential regression algorithm (Raghunathan, et al. , 2001) in IVEware or the “chained equations” approach (van Buuren et al., 1999; Royston, 2005; Carlin, et al., 2008) in Stata and R. Broadly described, each of these algorithms is based on an iterative algorithm. Each iteration (t=1,…,T) of the algorithm moves one-by-one through the sequence of variables in the imputation model, e.g. Y={Y_1,Y_2,Y_3,Y_4,Y_5} as illustrated in Figure 1.

	Variables				
Obs	Y1	Y2	Y3	Y4	Y5
1			?	?	
2	?	?		?	?
3	?		?		
4		?			
5	?			?	

Figure 1. Arbitrary Multivariate Missing Data Pattern

At each iteration and for each variable, there is a P-Step and I-Step. In the P-Step, the current (iteration *t*) values of the observed and imputed values for the imputation model variables are used to derive the predictive distribution of the missing values for the target variable. To model the conditional predictive distribution of individual Y_k, PROC MI uses the same regression or discriminant function methods available in PROC MI as for the monotone missing data patterns, (Berglund and Heeringa, 2014).

See Figure 2 (from the SAS/STAT PROC MI documentation) for a summary of all available imputation methods in SAS 9.4 and guidance on selection of an appropriate method.

Table 61.5: Imputation Methods in PROC MI

Pattern of Missingness	Type of Imputed Variable	Type of Covariates	Available Methods
Monotone	Continuous	Arbitrary	● Monotone regression ● Monotone predicted mean matching ● Monotone propensity score
Monotone	Classification (ordinal)	Arbitrary	● Monotone logistic regression
Monotone	Classification (nominal)	Arbitrary	● Monotone discriminant function
Arbitrary	Continuous	Continuous	● MCMC full-data imputation ● MCMC monotone-data imputation
Arbitrary	Continuous	Arbitrary	● FCS regression ● FCS predicted mean matching
Arbitrary	Classification (ordinal)	Arbitrary	● FCS logistic regression
Arbitrary	Classification (nominal)	Arbitrary	● FCS discriminant function

Figure 2. Table 61.5: Imputation Methods in PROC MI

MULTIPLE IMPUTATION OF COMPLEX SAMPLE DESIGN DATA

Complex surveys are comprised of data derived from sample designs that adjust for non-response and differing probabilities of selection. Complex samples differ from standard or simple random samples in that they assume independence of observations while complex samples do not. Most SAS procedures assume that data used is derived from a simple random sample and under-estimate variances when analyzing data from complex samples. Therefore, analysis of data from complex surveys should include methods of variance estimation that account for these sample design features (Kish, 1965 and Rust, 1985).

The SURVEY suite of procedures (PROC SURVEYSELECT, PROC SURVEYMEANS, PROC SURVEYFREQ, PROC SURVEYREG, PROC SURVEYLOGISTIC, and PROC SURVEYPHREG) allow the analyst to create samples and correctly analyze complex sample design data sets. However, another important consideration is how to correctly incorporate the complex sample design features and weights into the MI framework. Donald Rubin offered the following guidance on MI for complex samples: "Minimally, major clustering and stratification indicators and sample design weights (or estimated propensity scores of being in the sample) should be included in the imputation models. The possible lost

precision when including unimportant predictors is usually a small price to pay for the general validity of the resultant multiply imputed data base", (Rubin, 1996) .

To capture the complex sample design features and weight(s) in the imputation model, a recommended method is to create a categorical variable in the DATA STEP that is the combination of the stratum and cluster codes provided by the data producer. Then, use the combined strata and cluster variable along with the probability weight in the imputation model during MI Step 1. In Step 2, utilize the correct SAS SURVEY procedure with weights and design variables, i.e. single STRATA, CLUSTER, and WEIGHT variables to correctly analyze the imputed data sets and finally, use PROC MIANALYZE in MI Step 3 to combine results and produce valid inferences.

ANALYSIS APPLICATION

The analysis application is a detailed example that uses PROC MI with the FCS method to impute missing data on categorical variables with an arbitrary missing data pattern, analysis of imputed data sets using PROC SURVEYLOGISTIC, and analysis of results from MI Steps 1 and 2 using PROC MIANALYZE. Because SAS is of primary interest, a detailed discussion of code, output and interpretation is included in this section.

The application is then repeated using Stata, IVEware and R for direct comparison of results. For the replications, the focus is on the final pooled estimates rather than detailed explanations of the full syntax used. For more information on Stata, IVEware, or R, see their respective user manuals.

APPLICATION DATA SET

Data from the National Comorbidity Survey-Replication, a nationally representative sample based on a stratified, multi-stage area probability sample of the United States population (Kessler et al, 2004 and Heeringa, 1996) is used in the application. The NCS-R data set is based upon a complex sample design and contains variables representing the design features along with weights that adjust for non-response, differing probabilities of selection and post-stratification to a given population. See the project website at http://www.hcp.med.harvard.edu/ncs/ for more information.

VARIABLE LIST

The NCS-R data set is from the Part 2 of the survey (n=5,692) and includes a number of detailed questions about DSM-IV disorders and related issues such as treatment and impairment.

Variables used are in this application are as follows with variables with missing data highlighted in red:

- Sex (categorical, coded 0=FEMALE 1=MALE)
- Region (categorical, coded 1=NE 2=MW 3=SOUTH 4=WEST)
- Age (continuous, age in years)
- Str (continuous, strata representing complex sample design)
- Secu (categorical, cluster/PSU representing complex sample design)
- Finalp2wt (continuous, final part 2 weight)
- Racecat_ (categorical, coded 1=WHITE 2=HISPANIC 3=BLACK 4=OTHER)
- Educat (categorical, coded 1=0-11 YRS 2=12 YRS 3=13-15 YRS 4=16+ YRS, some missing data)
- MDE (categorical, coded 1=YES major depressive episode 0=NO MDE, some missing data)
- Str_Secu (categorical, combined Str and Secu variable)

EXAMINATION OF MISSING DATA

Prior to multiple imputation of missing data, an important preliminary step is to examine the data set for types of variables (continuous, categorical, count, etc.) that have missing data and the extent and pattern of missing data. Patterns of missing data can be broadly categorized as arbitrary, monotone, or matrix/file-matching, (see Figures 3-5 for graphic representations). Typically, identification of the missing data pattern helps drive the choice of imputation method and number of imputed data sets created during

MI Step 1. For more on the question of how many imputed data sets to create, see Table 61.7 of the PROC MI documentation.

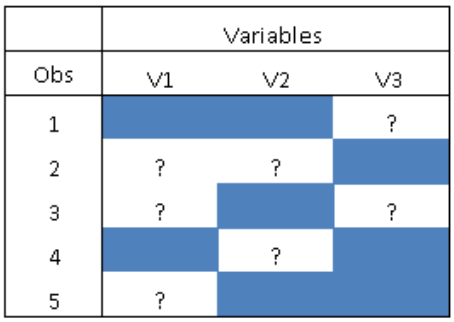

Figure 3. Arbitrary Missing Data

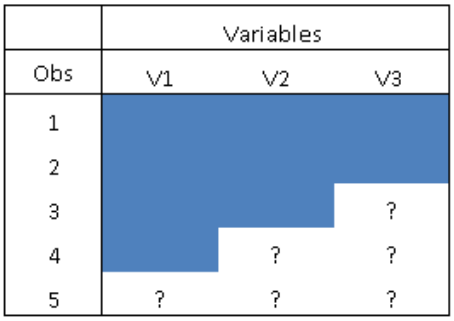

Figure 4. Monotone Missing Data

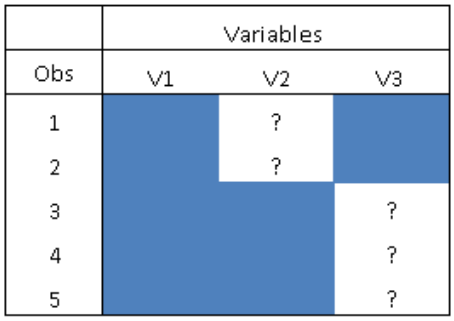

Figure 5. File-Matching or Matrix Missing Data

APPLICATION USING SAS 9.4

MI STEP 0 - EXPLORE MISSING DATA

The initial step, here called MI Step 0, explores the characteristics of missing data through use of PROC MI without imputation (NIMPUTE=0). PROC MI produces a Missing Data Pattern grid by default. The SAS code below reads in a temporary data set called NCSR2_1 and creates output in Figure 6 below:

```
proc mi nimpute=0 data=ncsr2_1;
run;
```

The MI Procedure

Model Information	
Data Set	D.NCSR2_1
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	0
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	556202000

Missing Data Patterns

Group	sex	region	age	str	secu	finalp2wt	racecat_	educat	mde	mde_imp	educat_imp	str_secu	Freq	Percent	Group Means: sex	region	age	str	secu	finalp2wt	racecat_	educat	mde	mde_imp	educat_imp	str_secu
1	X	X	X	X	X	X	X	X	X	X	X	X	5292	92.97	0.417989	2.579554	43.352419	26.376228	1.506236	0.992335	3.428193	2.649849	0.316515	0	0	265.268519
2	X	X	X	X	X	X	X	X	.	X	X	X	165	2.90	0.387879	2.454545	41.690909	26.957576	1.527273	1.084961	3.351515	2.703030	.	1.000000	0	271.103030
3	X	X	X	X	X	X	X	.	X	X	X	X	235	4.13	0.451064	2.557447	45.140426	26.029787	1.468085	1.112958	3.361702	.	0.293617	0	1.000000	261.765957

Figure 6. Missing Data Patterns, NCS-R Data Set

Based on Figure 6, the Model Information table contains basic information about the default imputation method used had there been an imputation (MCMC) along with other information related to the imputation process. Given that no imputation was actually performed, this information is not relevant to the process at this point.

The Missing Data Patterns table reveals an arbitrary missing data pattern with two variables that require imputation of missing data, EDUCAT and MDE. Both are classification variables (ordinal and binary, respectively) and even with re-ordering of the variables in the VAR statement, the missing data pattern would still be arbitrary. There are three distinct groups in the data set: 1. those with fully observed on all variables (92.97% of the 5,692 observations), 2. those missing on just Major Depressive Episode (MDE, 2.90%) and 3. those missing on only education in categories (EDUCAT, 4.13%). There are also two imputation flag variables constructed in the DATA STEP (code not shown here), MDE_IMP and EDUCAT_IMP. These flag variables are set equal to 1 for observations that are imputed and 0 otherwise.

MULTIPLE IMPUTATION STEP 1 - IMPUTE MISSING DATA

MI Step 1 uses PROC MI to impute missing data. The FCS imputation method is selected because it easily handles arbitrary missing data patterns with continuous or classification variables that need imputation.

The following code uses PROC MI to create a default 5 imputed data sets (NIMPUTE=5) using a SEED value (SEED=876) and creates a temporary output data set containing the (OUT=OUTFCS). In addition, the CLASS statement declares sex, region, race, education, MDE, and the combined strata and cluster variable as classification variables (CLASS SEX REGION RACECAT_ EDUCAT MDE STR_SECU). The FCS LOGISTIC statement requests the FCS logistic regression method with 40 burn-in iterations (NBITER=40) and model details (DETAILS) for each of the five imputation models used to impute MDE. Note that both variables to be imputed are binary (MDE) or ordinal (EDUCAT) and the LOGISTIC method is appropriate for both.

The VAR statement lists the variables to be used in the imputation models and omits the imputation flag variables as they do not have any scientific meaning in the imputation model. The final Part 2 NCS-R weight (FINALP2WT) and the combined strata and cluster variable (STR_SECU) are used as imputation model covariates to represent the complex sample design features and probability weight. Other model covariates include gender, US region, age at interview, race, and the imputed MDE (after MDE is imputed during the process). By default, PROC MI imputes the variables following the order in the VAR statement therefore, fully observed variables are listed first (SEX REGION AGE RACECAT_ STR_SECU FINALP2WT) followed by those with the least to the most missing data (MDE EDUCAT).

The output data set (OUTFCS) contains five imputed data sets stored in a "long" format along with a SAS generated variable called _IMPUTATION_ with values of 1-5 to identify each imputed data set. Therefore, the output data set contains 5*5,692=28,460 observations:

```
proc mi data=ncsr2_1 seed=876 nimpute=5 out=outfcs;
  class sex region racecat_ educat mde str_secu;
  fcs nbiter=40 logistic (mde/details) logistic (educat);
  var sex region age racecat_ str_secu finalp2wt mde educat;
run;
```

Model Information	
Data Set	WORK.NCSR2_1
Method	FCS
Number of Imputations	5
Number of Burn-in Iterations	40
Seed for random number generator	876

FCS Model Specification	
Method	Imputed Variables
Regression	age finalp2wt
Logistic Regression	mde educat
Discriminant Function	sex region racecat_ str_secu

Missing Data Patterns

Group	sex	region	age	racecat_	str_secu	finalp2wt	mde	educat	Freq	Percent	Group Means	
											age	finalp2wt
1	X	X	X	X	X	X	X	X	5292	92.97	43.352419	0.992335
2	X	X	X	X	X	X	X	.	235	4.13	45.140426	1.112958
3	X	X	X	X	X	X	.	X	165	2.90	41.690909	1.084961

Logistic Models for FCS Method

Imputed Variable	Effect	sex	region	racecat_	str_secu	educat	Imputation				
							1	2	3	4	5
mde	Intercept	.	.	.	.	.	1.314604	1.473147	1.573004	1.446349	1.288394
mde	sex	0	.	.	.	.	-0.191448	-0.142598	-0.147864	-0.138470	-0.145515
mde	region	.	1.000000	.	.	.	-1.290161	0.732726	-0.933802	-0.171140	-0.781971
mde	region	.	2.000000	.	.	.	0.913546	-0.265727	0.459304	0.391742	0.669216
mde	region	.	3.000000	.	.	.	0.067714	-0.416091	0.025426	-0.256251	0.178706
mde	age	.	.	.	.	.	0.045437	-0.030696	0.004936	0.036598	-0.004306
mde	racecat_	.	.	1.000000	.	.	0.082656	0.003019	-0.106662	-0.134692	-0.232575
mde	racecat_	.	.	2.000000	.	.	0.433958	0.242866	0.461350	0.259921	0.487857
mde	racecat_	.	.	3.000000	.	.	-0.315066	-0.010852	-0.093505	0.128259	-0.187522

Figure 7. Selected Output from the PROC MI FCS IMPUTATION

From Figure 7, the Model Information Table lists the FCS Method, Number of Burn-In Iterations and random number generator specified in the code. The FCS Model Specification lists the variables that require imputation along with Regression and Discriminant Function methods with associated variables for each method. Since the variables listed under the Regression and Discriminant Function methods are fully observed, the default methods are listed but **no** data is actually imputed for these variables.

The Missing Data Patterns grid displays the extent of missing data by group along with Group Means for just the continuous variables used in the imputation. Finally, the Logistic Models for FCS Method table (partial output presented) details the parameter estimates for each of five imputed data sets and for each level of the Effects. This level of detail is produced by the DETAILS option in the FCS statement and can be used as a diagnostic tool to evaluate the individual imputations. This output shows stable estimates across all 5 imputations of MDE for the effects sex, region, age, and race effects.

The code below illustrates use of PROC FREQ to produce unweighted cross-tabulations of observed MDE by imputed MDE (MDE*MDE_IMP), for each of the 5 imputed data sets. This type of informal diagnostic check permits evaluation of the observed v. imputed variable distributions with the aim of identifying possible problems in the imputation:

```
proc freq data=outfcs;
 tables _imputation_*mde*mde_imp / missing;
run;
```

Frequency
Percent
Row Pct
Col Pct

Table 1 of mde by mde_imp

Controlling for _Imputation_=1

mde(Major Depressive Episode)	mde_imp 0	mde_imp 1	Total
0	3783 66.46 97.15 68.45	111 1.95 2.85 67.27	3894 68.41
1	1744 30.64 97.00 31.55	54 0.95 3.00 32.73	1798 31.59
Total	5527 97.10	165 2.90	5692 100.00

Frequency
Percent
Row Pct
Col Pct

Table 2 of mde by mde_imp

Controlling for _Imputation_=2

mde(Major Depressive Episode)	mde_imp 0	mde_imp 1	Total
0	3783 66.46 97.30 68.45	105 1.84 2.70 63.64	3888 68.31
1	1744 30.64 96.67 31.55	60 1.05 3.33 36.36	1804 31.69
Total	5527 97.10	165 2.90	5692 100.00

Figure 8. Cross-Tabulations of Observed MDE and Imputed MDE by Imputation, For Data Sets 1 and 2 Only

The cross-tabulations in Figure 8 reveal how the values of imputed MDE differ across the imputed data sets (just results from imputations 1 and 2 are shown here). These slight differences reflect the expected random variability of the imputation process and present no evidence of problems in the imputation. For

example, for _IMPUTATION_=1, 67.27% of the respondents are imputed to MDE=0 and 32.73% are imputed to MDE=1. In comparison, the observed percentages are 68.45% (MDE=0) and 31.55% (MDE=1). The second imputed data set (_IMPUTATION_=2) shows similar differences between observed and imputed percentages. Other PROC MI diagnostic tools such as TRACE and AUTOCORRELATION plots are available for continuous variables, see the documentation for details and examples.

MULTIPLE IMPUTATION STEP 2 - ANALYZE IMPUTED DATA SETS USING PROC SURVEYLOGISTIC

With five imputed data sets produced in MI Step 1, MI Step 2 consists of analysis of the completed data sets using the SURVEY procedure of choice. The planned analysis for this example is a design-based logistic regression predicting the probability of having a diagnosis of lifetime Major Depressive Episode with gender, education, and US region covariates. Note that all of the variables to be used in the analysis plus additional covariates including the weight and complex sample design variables were included in the imputation models. In general, the imputation model should include, at the minimum, all analysis model variables plus additional meaningful covariates to enhance the imputations.

The code below reads the five imputed data sets stored in the OUTFCS data set (DATA=OUTFCS), the STRATA, CLUSTER, and WEIGHT statements represent the complex sample design and weights, and a CLASS statement is used with a REF option to declare classification variables and custom reference groups along with the PARAM=REF option to request reference group parameterization. Each design-based logistic regression is run separately within each imputed data set due to BY statement (BY _IMPUTATION_) and an output data set of parameter estimates and standard errors is created by ODS OUTPUT PARAMETERESTIMATES=OUTPARMS statement.

The PRINT procedure produces a listing report of the output data set from PROC SURVEYLOGISTIC. This data set will serve as input for PROC MIANALYZE in MI Step 3 and should contain, at a minimum, parameter estimates and variance information for univariate inference in MIANALYZE:

```
proc surveylogistic data=outfcs;
  strata str; cluster secu; weight finalp2wt;
  class sex (ref='0') educat (ref='1') region (ref='1') / param=ref;
  model mde (event='1') = sex educat region;
  by _imputation_;
  ods output parameterestimates = outparms;
run;
proc print data=outparms;
run;
```

Obs	_Imputation_	Variable	ClassVal0	DF	Estimate	StdErr	WaldChiSq	ProbChiSq
1	1	Intercept		1	-1.3040	0.1434	82.6946	<.0001
2	1	sex	1	1	-0.5106	0.0630	65.5785	<.0001
3	1	educat	2	1	0.0797	0.0983	0.6587	0.4170
4	1	educat	3	1	0.1895	0.1159	2.6742	0.1020
5	1	educat	4	1	0.1769	0.1145	2.3887	0.1222
6	1	region	2	1	0.00265	0.1320	0.0004	0.9840
7	1	region	3	1	-0.1698	0.1266	1.7985	0.1799
8	1	region	4	1	0.0623	0.1268	0.2413	0.6233
9	2	Intercept		1	-1.3062	0.1460	80.0374	<.0001
10	2	sex	1	1	-0.4840	0.0650	55.4803	<.0001
11	2	educat	2	1	0.0825	0.0955	0.7456	0.3879
12	2	educat	3	1	0.2312	0.1038	4.9605	0.0259
13	2	educat	4	1	0.1765	0.1120	2.4822	0.1151
14	2	region	2	1	0.00768	0.1407	0.0030	0.9565
15	2	region	3	1	-0.1848	0.1335	1.9150	0.1664
16	2	region	4	1	0.0334	0.1343	0.0618	0.8037

Figure 9. Partial Listing of the OUTPARMS Data Set from PROC SURVEYLOGISTIC, Imputation Data Sets 1 and 2

Figure 9 displays records from the OUTPARMS data set and includes a variable called _IMPUTATION_ with values of 1-5, Variable with the names of the model effects, CLASSVAL0 with the Class variable level, the degrees of freedom, estimated parameters and standard errors, and Wald Chi-Square values and associated *p* values.

MULTIPLE IMPUTATION STEP 3 - COMBINE RESULTS FROM MI STEPS 1 AND 2 AND GENERATE VALID INFERENCES USING PROC MIANALYZE

The third step of the MI process combines results from Steps 1 and 2 and generates valid inferences using PROC MIANALYZE. As a reminder, the output data set from Step 2 contains estimated weighted parameter estimates from the logistic regression predicting lifetime MDE with design-based standard errors from PROC SURVEYLOGISTIC. There are five sets of parameter estimates and standard errors that are combined by PROC MIANALYZE to reflect the variability of the imputation process along with the complex sample design features.

The following syntax executes PROC MIANALYZE and performs univariate inference. The PROC statement declares the OUTPARMS data set as a PARMS type of data set with classification variables read in with the CLASSVAL option (PARMS (CLASSVAR=CLASSVAL)=OUTPARMS). This option instructs SAS to read each CLASS variable's levels from the variable CLASSVAL0. The CLASS statement sets SEX, EDUCAT, and REGION as classification variables and omits the category specified in the SURVEYLOGISTIC code, the lowest category for each variable in the CLASS statement in this example. The MODELEFFECTS statement lists the model covariates, beginning with the intercept, in the order established in the previous step. Though this examples produces only univariate inferences, multivariate inference is possible in PROC MIANALYZE, see the SAS/STAT MIANALYZE documentation for details and examples:

```
proc mianalyze parms (classvar=classval)=outparms;
  class sex educat region;
  modeleffects intercept sex educat region;
run;
```

The SAS System

The MIANALYZE Procedure

Model Information	
PARMS Data Set	WORK.OUTPARMS
Number of Imputations	5

Variance Information

Parameter	sex	educat	region	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
				Between	Within	Total				
intercept				0.000838	0.021492	0.022498	2002.3	0.046787	0.045649	0.990953
sex	1			0.000181	0.004040	0.004257	1540.7	0.053689	0.052183	0.989671
educat		2		0.001012	0.010015	0.011229	342.21	0.121221	0.113282	0.977845
educat		3		0.000879	0.011435	0.012491	560.33	0.092288	0.087741	0.982754
educat		4		0.000414	0.012358	0.012855	2680.3	0.040184	0.039348	0.992192
region			2	0.000062849	0.019224	0.019299	261933	0.003923	0.003915	0.999218
region			3	0.000125	0.017197	0.017347	53719	0.008704	0.008666	0.998270
region			4	0.000456	0.017238	0.017785	4234.5	0.031709	0.031192	0.993800

Parameter Estimates

Parameter	sex	educat	region	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
intercept				-1.314977	0.149993	-1.60914	-1.02082	2002.3	-1.356894	-1.279889	0	-8.77	<.0001
sex	1			-0.498592	0.065245	-0.62657	-0.37061	1540.7	-0.511238	-0.483982	0	-7.64	<.0001
educat		2		0.090290	0.105967	-0.11814	0.29872	342.21	0.055653	0.141995	0	0.85	0.3948
educat		3		0.231947	0.111762	0.01242	0.45147	560.33	0.189506	0.271394	0	2.08	0.0384
educat		4		0.192461	0.113378	-0.02986	0.41478	2680.3	0.176522	0.223152	0	1.70	0.0897
region			2	0.005170	0.138923	-0.26712	0.27745	261933	-0.006845	0.014508	0	0.04	0.9703
region			3	-0.178517	0.131708	-0.43667	0.07963	53719	-0.195268	-0.169843	0	-1.36	0.1753
region			4	0.037567	0.133360	-0.22389	0.29902	4234.5	0.004407	0.062291	0	0.28	0.7782

Figure 10. Model Information, Variance Information and Multiple Imputation Logistic Regression Parameter Estimates for MDE: SAS Results from PROC MIANALYZE

Figure 10 includes selected tables from PROC MIANALYZE including Model Information, Variance Information, and Parameter Estimates. The Model Information lists the input OUTPARMS data set with 5 imputations.

The Variance Information table includes the between, within, and total variances for each parameter in the model. The table details the relative increase in variance due to missing data (range from 0.003 to 0.12) and Fraction Missing Information (range from 0.004 to 0.12) which reflects the impact of missing data among the variables used in the regression model. Based on five imputed data sets, Relative Efficiency is close to 1.0 for all effects, suggesting that five imputations are sufficient.

The Parameter Estimates represent averaged estimates with standard errors that are adjusted for both the complex sample design and the variability introduced by multiple imputation. Therefore, 95% confidence limits and *t* tests are based on the fully corrected standard errors.

These results suggest that men are significantly less likely than women to have an MDE diagnosis, those in higher education groups are more likely than the lowest educational group (0-11 years of education) to have MDE but only the group with 13-15 years of education is significant at the 0.05 alpha level. The results for US regions indicate that those in Midwest and West are more likely to have MDE as compared to those in the Northeast region while those in the South region are less likely than the Northeast region

to have MDE. None of the individual region predictors are significant and all results are interpreted while holding all other predictors in the model constant.

APPLICATION REPLICATIONS USING IVEWARE, STATA AND R SOFTWARE

IVEWARE

IVEware (Imputation and Variance Estimation Software) is a software designed to perform multiple imputation of missing data and subsequent analysis of data derived from complex sample designs. It uses the sequential regression method (also known as FCS or chained equations) to perform multiple imputation along with the Jackknife Repeated Replication (JRR) method for complex sample variance estimation.

In this demonstration, IVEware is used as a SAS-callable tool though it is also possible to run the software as a standalone version, see iveware.org for more information and downloads. The software performs imputation with the *%IMPUTE* macro and regression analysis with correct variance estimation with the *%REGRESS* macro and descriptive analysis with the *%DESCRIBE* macro. The macros are run from the regular (not enhanced) SAS program editor and SAS programmers with a basic understanding of how to invoke macros can execute the IVEware program without the need to learn a new language.

The following syntax reads in the SAS data set named NCSR2_1 and performs preliminary recodes in the DATA STEP prior to multiple imputation. The *%IMPUTE* macro imputes missing data using the sequential regression method and creates 5 multiples or imputed data sets. This macro call uses a number of additional statements to control the imputation. For example, the default variable type is set to CONTINUOUS while classification variables are declared as CATEGORICAL and the remaining variables in the data set are declared as TRANSFER, meaning these variables are retained but not used in the imputation models. A SEED value is used to ensure future replication of the results and MULTIPLES is set to 5 to produce five imputed data sets:

```
data app4;
 set ncsr2_1;
* Recode the dependent variable to make highest category (no) the omitted;
 if mde=0 then mde_r=2; else if mde=1 then mde_r=1;  else mde_r=.;
run ;

%impute (name=app4, setup=new, dir=. );
datain  app4;
dataout app4_imp;
default continuous;
categorical sex region racecat_ educat mde_r str_secu;
transfer sampleid mde_imp educat_imp str secu mde;
multiples 5;
seed 876;
run;
```

After imputation, the *%PUTDATA* macro outputs five temporary SAS data sets (IMP1-IMP5) to be used as input to the *%REGRESS* macro. *%REGRESS* performs design-based logistic regression with the JRR variance estimation method while the LINK LOGISTIC statement requests a logistic regression using the outcome variable MDE_R. This recoded variable predicts the probability of having MDE (coded as 1) while no MDE (coded as 2) serves as the omitted category. Since IVEware omits the highest category of any categorical variable, the reference categories differ from SAS, Stata, and R. An alternative is to use indicator variables representing each level of the categorical variables and omit the lowest category to match the other programs (not shown here).

Use of the complex sample design variables and weight with the five imputed input data sets produce regression results that incorporate the imputation variability and complex sample design features:

```
* use %putdata to produce 5 separate data sets for correct MI estimation;
%putdata(name=app4,dir=.,  mult=1,dataout=imp1 );
%putdata(name=app4,dir=.,  mult=2,dataout=imp2 );
%putdata(name=app4,dir=.,  mult=3,dataout=imp3 );
%putdata(name=app4,dir=.,  mult=4,dataout=imp4 );
%putdata(name=app4,dir=.,  mult=5,dataout=imp5 );

%regress (name=app4_2, setup=new, dir=. );
datain  imp1 imp2 imp3 imp4 imp5;
stratum str;
cluster secu;
weight  finalp2wt;
categorical sex educat region;
predictor sex educat region;
dependent mde_r;
link logistic;
run ;
```

```
All imputations
Valid cases              5692
Sum weights        5692.000487

Degr freedom        179.406391

-2 LogLike         5487.725767

Variable              Estimate        Std Error        Wald test       Prob > Chi
Intercept          -1.6017202        0.0937348         291.99244          0.00000
sex                 0.5050432        0.0639964          62.27970          0.00000
educat.1           -0.1806331        0.1147336           2.47864          0.11540
educat.2           -0.0954590        0.0895300           1.13683          0.28632
educat.3            0.0602163        0.0795369           0.57318          0.44900
region.1           -0.0341966        0.1439643           0.05642          0.81224
region.2           -0.0374815        0.1034633           0.13124          0.71715
region.3           -0.2036425        0.0916866           4.93316          0.02635

Variable                 Odds            95% Confidence Interval
                        Ratio            Lower             Upper
Intercept
sex                 1.6570571        1.4604735         1.8801014
educat.1            0.8347416        0.6656214         1.0468315
educat.2            0.9089556        0.7617580         1.0845969
educat.3            1.0620662        0.9077993         1.2425485
region.1            0.9663815        0.7274006         1.2838775
region.2            0.9632122        0.7853363         1.1813764
region.3            0.8157540        0.6807464         0.9775367

Variable               Design           SRS            % Diff
                       Effect         Estimate         SRS v Est
Intercept             1.21094       -0.9769619         -39.00545
sex                   1.12449        0.5185654           2.67744
educat.1              1.39625       -0.1475597         -18.30971
educat.2              1.15261       -0.0470697         -50.69118
educat.3              1.04220        0.0265876         -55.84643
region.1              2.46956       -0.1698988         396.83027
region.2              1.56901       -0.0118203         -68.46355
region.3              1.30510       -0.1530469         -24.84526
```

Figure 11. Multiple Imputation Logistic Regression of MDE: Results From IVEware

Figure 11 includes parameter estimates with JRR based variance estimates, Wald tests, Odds Ratios with 95% Confidence Limits, Design Effects, SRS estimates, and the percentage difference between the SRS and design-based Estimates. The variances are adjusted for the variability due to the MI process as well

as the complex sample design features and weight. Interpretation of the results is included in a later section where the results from all four software packages are contrasted.

STATA

The next replication uses Stata (v13.1) to impute missing data and analyze imputed data sets while taking the complex sample design features and weight into account. Stata offers a number of multiple imputation and survey commands within the *mi* and *svy* suite of commands, see the Stata documentation for details.

The command syntax below reads the input data set, sets up the multiple imputation by registering imputed and regular (not imputed) variables, imputes missing data using the "chained equations" method, sets the survey variables and then performs an MI and design-based analysis using the *mi estimate:svy: logit* command:

```
* read data set into memory
use "ncsr2_v12.dta", clear

* set up mi data and register variables
mi set mlong
mi register imputed mde educat
mi register regular sex region racecat_ age finalp2wt str secu str_secu

* impute missing data using chained logit, ologit commands
mi impute chained (logit) mde (ologit) educat=i.sex i.region ///
 i.racecat_ age finalp2wt i.str_secu , add(5) rseed(2012)

* set survey variables within the mi suite of commands
mi svyset secu [pweight=finalp2wt], strata(str)

* run mi estimate: svy logit regression
mi estimate: svy: logit mde i.sex i.region i.educat
```

```
Multiple-imputation estimates                   Imputations        =         5
Survey: Logistic regression                     Number of obs      =      5692
Number of strata  =        42                   Population size    = 5692.0005
Number of PSUs    =        84
                                                Average RVI        =    0.0464
                                                Largest FMI        =    0.0933
                                                Complete DF        =        42
DF adjustment:   Small sample                   DF:     min        =     34.46
                                                        avg        =     38.43
                                                        max        =     39.79
Model F test:       Equal FMI                   F(   7,   39.7)    =     16.02
Within VCE type:   Linearized                   Prob > F           =    0.0000
```

```
------------------------------------------------------------------------------
         mde |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.sex |  -.5111926   .0638813    -8.00   0.000    -.6403719   -.3820132
             |
      region |
           2 |   .0065904   .1408085     0.05   0.963    -.2783006    .2914814
           3 |  -.1669361   .1330025    -1.26   0.217    -.4358414    .1019692
           4 |    .039644   .1319837     0.30   0.765    -.2271491    .3064372
             |
      educat |
           2 |    .094891   .1074464     0.88   0.383    -.1228383    .3126204
           3 |   .2324122   .1124072     2.07   0.046      .004085    .4607393
           4 |   .1770819   .1114483     1.59   0.120    -.0482573    .4024212
             |
       _cons |  -1.311143   .1484784    -8.83   0.000    -1.611389   -1.010897
```

Figure 12. Multiple Imputation Logistic Regression of MDE: Results From Stata

Figure 12 includes information about the number of imputations contained in the multiply imputed data set (5), variance information such as Average RVI (Relative Variance Increase) and Largest FMI (Fraction Missing Information), Degrees of Freedom (complete and Small Sample adjusted), and an *F* test for the model. In addition, parameter estimates, standard errors, *t* tests and *p* values, and 95% confidence intervals that account for the MI process and the complex sample design are presented. As with IVEware and R, interpretation of results is done in the last section of this paper.

R

The final replication uses R v3.0.1 with the *mice, mitools, foreign*, and *survey* packages to impute missing data using chained equations and analyze imputed data sets with *mitools* commands and the *svyglm* command for design-based logistic regression that also accounts for the multiple imputation variability.

The following code loads the needed R packages, reads in a Stata format data set and translates for use in R, creates factor variables for use in the multiple imputation and subsequent analyses, imputes missing data using the *mice* (multiple imputation by chained equations) command, and converts the output MI data to a format acceptable for use with the *mitools* package. Then, the syntax sets the complex sample design variables and weight and executes the *svyglm* command with the correct family option to perform a multiple imputation, design-based logistic regression while using the five imputed data sets with the *MIcombine* command:

```
# load packages using library command
library(foreign)
library(mi)
library(mice)

# read Stata format data set into R
a <- read.dta("C:/ncsr2_v12.dta" )
summary(a)

# create factor variables
a$sex <- factor(a$sex)
a$educat <- factor (a$educat)
a$region <- factor(a$region)
a$str_secu <- factor(a$str_secu)

# obtain information about missing data
```

```
inf <-mi.info(a)
# print info about missing data
inf

# use mice to impute and pool
library(mice)
imp <- mice(a,n.imp=5,seed=1934)
summary(imp)

# convert mids to data useable for work in mitools
library(mitools)
mydata <- imputationList(lapply(1:5, complete, x=imp))
summary(mydata)

# set survey design
library(survey)
des <- svydesign(id=~secu, strat=~str, weight=~finalp2wt, data=(mydata), nest=TRUE)
summary(des)

# run design based model with svyglm using 5 imputed data sets contained in des (from
mydata)
fit2 <- with (des, svyglm (mde ~ sex + educat + region, family=quasibinomial))
summary(MIcombine(fit2))
```

```
Multiple imputation results:
      with(des, svyglm(mde ~ sex + educat + region, family = quasibinomial))
      MIcombine.default(fit2)
                results         se      (lower     upper) missInfo
(Intercept) -1.48749336 0.14096000 -1.76381159 -1.2111751      2 %
sex2        -0.39672207 0.08145831 -0.55638193 -0.2370622      1 %
educat2      0.11943740 0.10932972 -0.09530298  0.3341778      9 %
educat3      0.25810927 0.11257402  0.03738999  0.4788286      3 %
educat4      0.18291288 0.12037457 -0.05308465  0.4189104      3 %
region2      0.08537139 0.13342511 -0.17615071  0.3468935      1 %
region3     -0.06108228 0.13851353 -0.33256725  0.2104027      1 %
region4      0.10233875 0.13072914 -0.15390103  0.3585785      1 %
```

Figure 13. Multiple Imputation Logistic Regression of MDE: Results From R

The default output from the R *svyglm* and *MIcombine* commands includes parameter estimates, standard errors (from the TSL variance estimation method), 95% confidence limits, and percentage of missing information. Interpretation of the R results is presented in the next section.

COMPARISON OF MULTIPLE IMPUTATION LOGISTIC REGRESSION RESULTS FROM SAS, STATA, IVEWARE, AND R

The results from all four software tools are presented in Table 1. Each software uses MI logistic regression to predict the probability of having a diagnosis of lifetime Major Depressive Episode with gender, education, and US region covariates. The five multiply imputed data sets are analyzed with logistic regression procedures that account for both the complex sample design and the variability introduced by multiple imputation.

Despite differences in omitted categories, methods of design-based variance estimation, imputation models, and differing seed values, there are few differences in the overall conclusions.

Outcome is Major Depressive Episode (LT)	SAS		Stata		IVEware		R	
Parameter	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
Male	-0.50*	0.07	-.52*	0.06	-	-	-.40*	0.08
Female	-	-	-	-	0.51*	0.06	-	-
0-11 Years Education	-	-	-	-	-0.18	0.11	-	-
12 Years Education	0.09	0.11	0.09	0.11	-0.10	0.09	0.12	0.11
13-15 Years Education	0.23*	0.11	0.23*	0.11	0.06	0.08	0.26*	0.11
16+ Years Education	0.19	0.11	0.18	0.11	-	-	0.18	0.12
Northeast Region	-	-	-	-	-0.03	0.14	-	-
Midwest Region	0.01	0.14	0.01	0.14	-0.04	0.10	0.09	0.13
South Region	-0.18	0.13	-0.17	0.13	-0.20*	0.09	-0.06	0.14
West Region	0.04	0.13	0.04	0.13	-	-	0.10	0.13

* Significant at the alpha-0.05 level.

Table 1. Comparison of Multiple Imputation Logistic Regression of MDE: Results from SAS, Stata, IVEware, and R

Based on Table 1, the SAS, Stata, and R results suggest that compared to females, men are significantly less likely to have MDE, those with some college are significantly more likely than those with 0-11 years of education to experience MDE and compared to the Northeast region, none of the other regions of the United States are significant at the alpha=0.05 level. All of these interpretations are population estimates of these relationships when holding the other predictor variables in the model at fixed values.

Though IVEware uses different reference groups than SAS, Stata, and R, as expected, the overall conclusions do not change. For example, compared to those with 16+ years of education, those in the two lowest educational groups (0-11 years and 12 years of education) are less likely to be diagnosed with MDE while those with some college are more likely to have MDE but none are significant. Women are significantly more likely than men to have MDE, and those living in the Northeast, Midwest, or South regions are less likely to have MDE as compared to those in the West, holding the other predictor variables in the model at fixed values.

CONCLUSION

This paper presents a detailed application of multiple imputation of missing data from a complex sample design data set using the PROC MI Fully Conditional Specification (FCS) method with subsequent analysis using PROC SURVEYLOGISTIC and PROC MIANALYZE. The FCS method is an excellent option for imputation of continuous and classification variables with an arbitrary missing data pattern.

The application is replicated using Stata, IVEware, and R with an equivalent imputation method while accounting for the complex sample design features and weights. The general comparisons of the regression results reveal no major differences in overall conclusions.

REFERENCES

Allison, Paul D., “Missing Data”, Sage Publications 2001.

Berglund and Heeringa (2014), “Multiple Imputation of Missing Data Using SAS”, SAS Publishing

Carlin, J.B., Galati, J.C., and Royston, P., A new framework for managing and analyzing multiply imputed data in Stata, *The Stata Journal*, 8(1), 49-67, 2008.

Heeringa, S., “Imputation Module Notes from Analysis of Complex Sample Data,” Institute for Social Research, University of Michigan Summer Institute Training Program.

Heitjan, Daniel F. 1997. “Annotation: What can be done about missing data? Approaches to imputation.” *American Journal of Public Health* 87: 548–550.

Horton, N.J. and Lipsitz, S.R. 2001. “Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables.” *Journal of the American Statistical Association* 55: 244–254.

Kish, L., *Survey Sampling*, John Wiley & Sons, New York, 1965.

Raghunathan, T.E., Lepkowski, J., Van Hoewyk, J., and Solenberger, P. 2001.
“A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models.” Survey Methodology: 27, pages 85-95.

Reiter, J.P., Raghunathan, T.E., and Kinney, V. 2006.
“The importance of modeling the sampling design in multiple imputation for missing data.” *Survey Methodology* 32.2: 143–150.

Royston, P., Multiple imputation of missing values, *Stata Technical Journal*, 5(4), 527–536, 2005.

Rubin, D.B., Multiple imputation after 18+ years, *Journal of the American Statistical Association*, 91(434), 473-489, 1996.

Rubin, D.B., *Multiple Imputation for Nonresponse in Surveys,* John Wiley & Sons, New York, 1987.

Rust, K. (1985), “Variance Estimation for Complex Estimators in Sample Surveys,” *Journal of Official Statistics*, 1, 381–397.

Van Buuren, S. and Oudshoorn, C.G.M. (1999), Flexible multivariate imputation by MICE, Leiden: TNO Preventie en Gezondheid, TNO/VGZ/PG 99.054.

Van Buuren, S. (2012), Flexible Imputation of Missing Data. Chapman & Hall/CRC, Boca Raton, FL.

CONTACT INFORMATION

Your comments and questions are valued and encouraged!

Contact the author at:

Patricia A. Berglund

University of Michigan-Institute for Social Research

pberg@umich.edu