How to Combine Inferences from Multiply-Imputed Data Sets: SAS and STATA Examples

Bo Klauth*

February 2, 2023

1. Introduction

This paper presents examples for producing estimates using Census-Enhanced Health and Retirement Study (CenHRS) data. The CenHRS crosswalk is constructed using a mix of both deterministic and probabilistic matching of HRS survey respondents to Census business data. Multiple imputation (MI) is necessary to conduct valid statistical inference with CenHRS data because it allows users to properly account for the additional variability induced by probabilistic matching. As such, the CenHRS crosswalk is designed with ten imputations for each HRS record. Because each imputation represents a single completed data set, there are in effect ten data sets, which have been stacked into a single file. Users should not attempt to calculate point estimates, standard errors, or confidence intervals for only one set of imputations or by treating the stacked file as a single data set without using appropriate MI analysis methods. For more information, see Rubin (1987).

This paper presents examples showing how these estimates are combined using SAS and STATA. SAS examples include how to combine means using the SAS PROC UNIVARIATE procedure and regression model coefficients from the SAS PROC REG, PROC GLM, or PROC MIXED procedures using the SAS PROC MIANALYZE procedure. STATA examples include how to combine means and regression model coefficients using `mi estimate: mean` and `mi estimate: regress`, respectively.

2. Example Data

For our examples, we create a fictitious longitudinal data set simulating the structure of CenHRS data containing the variable `year` (taking on values 2010, 2011, and 2012), `pid` (a person identifier), two analysis variables `firm_size` (firm size) and `hr_wage` (hourly wages), and `Replicate` (an implicate identifier that runs from 1 to 10) to demonstrate various ways users can combine the estimates from multiply-imputed data sets. Please find the data sets (SAS: `data_mi_2010_2012.sas7bdat`; STATA: `data_mi_2010_2012.dta`) accompanying this document from the CenHRS website: https://cenhrs.isr.umich.edu/documentation.

To demonstrate how to use SAS and STATA to compute estimates using the multiply imputed data, we provide examples to compute summary statistics for the analysis variables as well as to compute regression model coefficients for the relationship between firm size as the independent variable and hourly wages as the dependent variable.

*Survey Research Center, Institute for Social Research, University of Michigan. Please contact CenHRSinfo@umich.edu for questions related to this paper or the CenHRS project.

3. SAS Examples

We first demonstrate how users can combine means using the SAS PROC UNIVARIATE procedure and regression model coefficients from the SAS PROC REG, PROC GLM, or PROC MIXED procedures using the SAS PROC MIANALYZE procedure.

3.1. Modifying the Data to Use SAS

For our SAS code examples, we assume that users have downloaded the example data file (`data_mi_2010_2012.sas7bdat`) and stored it in a local directory, which we refer to as `C:\mi`.

The PROC MIANALYZE procedure is used to combine estimates from multiple imputations. To use the PROC MIANALYZE procedure, users must rename the imputation variable ***Replicate*** to ***_Imputation_***, which is the variable name required by the procedure. In the code that follows, users can specify the location of the data using LIBNAME, load the data set, and rename the imputation variable. Then, PROC PRINT prints ten observations to view. See Output 01.

```
*Create Library;
libname mi "C:\mi";

Title "Analyzing multiply-imputed Data Sets";

* Modify some variables;
data mydata;
    set mi.data_mi_2010_2012;
    rename Replicate = _Imputation_;
run;


proc print data = mydata (obs = 10);
    title2 "Output 01: multiply-imputed Data (10 Observations)";
run;
```

Analyzing Multiply-Imputed Data Sets
Output 01: Multiply-Imputed Data (10 Observations)

Obs	firm_size	hr_wage	pid	Replicate	year
1	10690	22.22	388	1	2010
2	16068	29.29	115	1	2010
3	6374	22.22	13	1	2010
4	5527	12.12	348	1	2010
5	2467	21.21	154	1	2010
6	2552	9.09	319	1	2010
7	11809	48.48	399	1	2010
8	4712	24.24	343	1	2010
9	19259	42.42	34	1	2010
10	23209	26.26	197	1	2010

3.2. Estimating Means and Standard Errors

Users can use PROC UNIVARIATE to compute the means and standard errors for the two variables. Note that users need to include the line "`by _Imputation_;`" in the procedure. Users can also have SAS output the results to a table called *outuni*. PROC PRINT prints the output for 2010, which includes 10 imputations. The results are shown in Output 02. Users should note that not all analysis variables vary across imputations as shown in Output 02. For example, if a data set containing HRS variables (e.g., hourly wages), which only have one imputation, is merged with a data file with 10 imputations, the between variance for the HRS variables across the 10 imputations in the merged data set should be zero.

```
* Sort data by year and _Imputation_;
proc sort data = mydata;
    by _Imputation_;
run;

* Calculate means from multiply-imputed Data Sets;
proc univariate data = mydata;
    var hr_wage firm_size;
    output out = outuni mean = hr_wage firm_size
    stderr = se_hr_wage se_firm_size;
    by _Imputation_;
    where year = 2010;
run;

* Print the output;
proc print data = outuni (obs = 10);
    title2 'Output 02: UNIVARIATE Means and Standard Errors (10 Imputations)';
run;
```

Analyzing Multiply-Imputed Data Sets
Output 02: UNIVARIATE Means and Standard Errors (10 Imputations)

Obs	year	_Imputation_	hr_wage	firm_size	se_hr_wage	se_firm_size
1	2010	1	25.2177	14772.05	0.52544	481.407
2	2010	2	25.2177	14538.94	0.52544	450.074
3	2010	3	25.2177	14672.19	0.52544	453.076
4	2010	4	25.2177	14913.87	0.52544	466.780
5	2010	5	25.2177	14528.81	0.52544	458.591
6	2010	6	25.2177	14607.10	0.52544	450.552
7	2010	7	25.2177	14636.54	0.52544	453.645
8	2010	8	25.2177	14298.77	0.52544	456.432
9	2010	9	25.2177	14645.93	0.52544	469.530
10	2010	10	25.2177	14066.64	0.52544	445.643

Because the output *outuni* contains 10 means and standard errors derived from the 10 imputations, users will need to use PROC MIANALYZE to combine those means and standard errors.

```
* Combine the means;
proc mianalyze data = outuni;
    modeleffects hr_wage firm_size;
    stderr se_hr_wage se_firm_size;
    by year;
    title2 "Output 03: Combine Means from multiply-imputed Data Sets for 2010";
run;
```

PROC MIANALYZE produces results as shown in Output 03. Because the between variance for hourly wages equals zero, relevant statistics cannot be computed and are set to missing by SAS. The last table shows the combined estimates (means and standard errors) and the p values (`Pr > |t|`) for the firm size and hourly wages variables.

Analyzing Multiply-Imputed Data Sets
Output 03: Combine Means from Multiply-Imputed Data Sets

The MIANALYZE Procedure

year=2010

Model Information	
Data Set	WORK.OUTUNI
Number of Imputations	10

Variance Information (10 Imputations)

Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
hr_wage	0	0.276091	0.276091	.	0	.	.
firm_size	56738	210396	272809	171.96	0.296641	0.237593	0.976792

Parameter Estimates (10 Imputations)

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
hr_wage	25.217680	0.525444	.	.	.	25.217680	25.217680	0	.	.
firm_size	14568	522.310835	13537.12	15599.05	171.96	14067	14914	0	27.89	<.0001

3.3. Estimating Model Coefficients Using PROC REG

Using the regression model specified in Section 2 estimating the relationship of firm size on hourly wages, users can analyze the data using PROC REG by including the "`by _Imputation_;`" statement. Here, we estimate model coefficients for 2010.

Users also need to output the coefficient table for PROC MIANALYZE. This table is named *`outreg`* in this example.

```
title2;
* Sort data by imputation;
proc sort data = mydata;
  by _Imputation_;
run;

*Using PROC REG;
proc reg data = mydata outest = outreg covout ;
  model hr_wage = firm_size;
  by _Imputation_;
  where year = 2010;
run;

* Print regression output;
proc print data = outreg (obs = 10);
  title2 "Output 04: PROC REG Output from multiply-imputed data sets";
run;
```

The output is shown in Output 04.

Analyzing Multiply-Imputed Data Sets
Output 04: PROC REG Output from Multiply-Imputed Data Sets

Obs	_Imputation_	_MODEL_	_TYPE_	_NAME_	_DEPVAR_	_RMSE_	Intercept	firm_size	hr_wage
1	1	MODEL1	PARMS		hr_wage	10.8268	18.9200	0.000426326	-1
2	1	MODEL1	COV	Intercept	hr_wage	10.8268	0.6768	-.000029946	.
3	1	MODEL1	COV	firm_size	hr_wage	10.8268	-0.0000	0.000000002	.
4	2	MODEL1	PARMS		hr_wage	10.8517	18.6733	0.000450130	-1
5	2	MODEL1	COV	Intercept	hr_wage	10.8517	0.7280	-.000033876	.
6	2	MODEL1	COV	firm_size	hr_wage	10.8517	-0.0000	0.000000002	.
7	3	MODEL1	PARMS		hr_wage	10.8090	18.5109	0.000457109	-1
8	3	MODEL1	COV	Intercept	hr_wage	10.8090	0.7247	-.000033470	.
9	3	MODEL1	COV	firm_size	hr_wage	10.8090	-0.0000	0.000000002	.
10	4	MODEL1	PARMS		hr_wage	10.8061	18.5911	0.000444322	-1

Next, users can use PROC MIANALYZE to combine model coefficients. The results are shown in Output 05. The last table shows the combined estimates (coefficients and standard errors) and the p values (`Pr > |t|`) for the firm size and hourly wage variables.

```
* Combine regression coefficients from multiply-imputed data sets;
proc mianalyze data = outreg;
    modeleffects Intercept firm_size;
    title2 "Output 05: Combine Model Coefficients from PROC REG";
run;
```

Analyzing Multiply-Imputed Data Sets
Output 05: Combine Model Coefficients from PROC REG

The MIANALYZE Procedure

Model Information	
Data Set	WORK.OUTREG
Number of Imputations	10

Variance Information (10 Imputations)

Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
Intercept	0.018016	0.709118	0.728936	12176	0.027947	0.027347	0.997273
firm_size	8.313198E-11	2.2379772E-9	2.3294224E-9	5840.1	0.040861	0.039585	0.996057

Parameter Estimates (10 Imputations)

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	18.698838	0.853777	17.02530	20.37238	12176	18.510893	18.919971	0	21.90	<.0001
firm_size	0.000448	0.000048264	0.00035	0.00054	5840.1	0.000426	0.000457	0	9.27	<.0001

3.4. Estimating Model Coefficients Using PROC GLM

The code that follows shows how users can use PROC GLM to run the regression model and produce necessary output tables for PROC MIANALYZE. The "`by _Imputation_;`" statement is needed. In addition, users need to obtain the model coefficients and the inverse matrix produced by PROC GLM.

```
title2;
* Using PROC GLM;
proc glm data = mydata;
    model hr_wage = firm_size/inverse;
    by _Imputation_;
    ods output ParameterEstimates = glmparms
    InvXPX = glmxpxi;
    where year = 2010;
run;

proc print data = glmparms (obs = 8);
    var _Imputation_ Parameter Estimate StdErr;
    title2 'Output 06: GLM Model Coefficients';
run;

proc print data = glmxpxi (obs = 12);
    var _Imputation_ Parameter Intercept firm_size;
    title2 'Output 07: GLM X''X Inverse Matrices';
run;
```

The model coefficients and inverse matrix are shown in Outputs 06 and 07, respectively.

Analyzing Multiply-Imputed Data Sets
Output 06: GLM Model Coefficients

Obs	_Imputation_	Parameter	Estimate	StdErr
1	1	Intercept	18.91997105	0.82268388
2	1	firm_size	0.00042633	0.00004502
3	2	Intercept	18.67327023	0.85325351
4	2	firm_size	0.00045013	0.00004827
5	3	Intercept	18.51089317	0.85131635
6	3	firm_size	0.00045711	0.00004776
7	4	Intercept	18.59112269	0.84339896
8	4	firm_size	0.00044432	0.00004635

Analyzing Multiply-Imputed Data Sets
Output 07: GLM X'X Inverse Matrices

Obs	_Imputation_	Parameter	Intercept	firm_size
1	1	Intercept	0.0057738552	-2.554727E-7
2	1	firm_size	-2.554727E-7	1.729433E-11
3	1	hr_wage	18.919971053	0.0004263261
4	2	Intercept	0.0061824188	-2.876702E-7
5	2	firm_size	-2.876702E-7	1.978619E-11
6	2	hr_wage	18.673270233	0.0004501298
7	3	Intercept	0.0062031777	-2.864724E-7
8	3	firm_size	-2.864724E-7	1.952485E-11
9	3	hr_wage	18.510893166	0.0004571087
10	4	Intercept	0.006091533	-2.743442E-7
11	4	firm_size	-2.743442E-7	1.839524E-11
12	4	hr_wage	18.591122685	0.0004443218

Finally, users can use PROC MIANALYZE to combine the model coefficients. The results are shown in Output 08. The last table shows the combined estimates (coefficients and standard errors) and the p values (`Pr > |t|`) for firm size and hourly wages.

```
* Combine regression coefficients from multiply-imputed data sets;
proc mianalyze parms = glmparms xpxi = glmxpxi;
   modeleffects Intercept firm_size;
```

```
   title2 "Output 08: Combine Regression Coefficients from PROC GLM";
run;
```

Analyzing Multiply-Imputed Data Sets
Output 08: Combine Regression Coefficients from PROC GLM

The MIANALYZE Procedure

Model Information	
PARMS Data Set	WORK.GLMPARMS
XPXI Data Set	WORK.GLMXPXI
Number of Imputations	10

Variance Information (10 Imputations)

Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
Intercept	0.018016	0.709118	0.728936	12176	0.027947	0.027347	0.997273
firm_size	8.313198E-11	2.2379772E-9	2.3294224E-9	5840.1	0.040861	0.039585	0.996057

Parameter Estimates (10 Imputations)

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	18.698838	0.853777	17.02530	20.37238	12176	18.510893	18.919971	0	21.90	<.0001
firm_size	0.000448	0.000048264	0.00035	0.00054	5840.1	0.000426	0.000457	0	9.27	<.0001

3.5. Estimating Model Coefficients Using PROC MIXED

If users wish to use PROC MIXED to perform a regression analysis, the code that follows shows how to output the model coefficients as an input table for PROC MIANALYZE. The "`by _Imputation_;`" statement is needed. The model coefficients are shown in Output 09.

```
title2;
* Using PROC MIXED;
proc mixed data = mydata;
    * put a class statement here if applicable;
    model hr_wage =  firm_size/ solution;
    by _Imputation_;
  where year = 2010;
    ods output SolutionF = mxparms;
run;

* Print the output from proc mixed;
proc print data = mxparms (obs = 8);
    title2 'Output 09: MIXED Model Coefficients';
run;
```

Analyzing Multiply-Imputed Data Sets
Output 09: MIXED Model Coefficients

Obs	_Imputation_	Effect	Estimate	StdErr	DF	tValue	Probt
1	1	Intercept	18.9200	0.8227	498	23.00	<.0001
2	1	firm_size	0.000426	0.000045	498	9.47	<.0001
3	2	Intercept	18.6733	0.8533	498	21.88	<.0001
4	2	firm_size	0.000450	0.000048	498	9.33	<.0001
5	3	Intercept	18.5109	0.8513	498	21.74	<.0001
6	3	firm_size	0.000457	0.000048	498	9.57	<.0001
7	4	Intercept	18.5911	0.8434	498	22.04	<.0001
8	4	firm_size	0.000444	0.000046	498	9.59	<.0001

Finally, users can use PROC MIANALYZE to combine the model coefficients from multiply-imputed data sets. The final combined coefficients are shown in Output 10 in the last table.

```
* Combine model coefficients from multiply-imputed data sets;
proc mianalyze parms(classvar = full) = mxparms;
   * put a class statement here if applicable;
   modeleffects Intercept firm_size;
   title2 "Output 10: Combine Model Coefficients from PROC MIXED";
run;
```

Analyzing Multiply-Imputed Data Sets
Output 10: Combine Model Coefficients from PROC MIXED

The MIANALYZE Procedure

Model Information	
PARMS Data Set	WORK.MXPARMS
Number of Imputations	10

Variance Information (10 Imputations)							
	Variance				Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
Parameter	Between	Within	Total	DF			
Intercept	0.018016	0.709118	0.728936	12176	0.027947	0.027347	0.997273
firm_size	8.313198E-11	2.2379772E-9	2.3294224E-9	5840.1	0.040861	0.039585	0.996057

Parameter Estimates (10 Imputations)										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	18.698838	0.853777	17.02530	20.37238	12176	18.510893	18.919971	0	21.90	<.0001
firm_size	0.000448	0.000048264	0.00035	0.00054	5840.1	0.000426	0.000457	0	9.27	<.0001

4. STATA Examples

We next demonstrate how users can combine means and regression model coefficients using the STATA `mi estimate: mean` and `mi estimate: regress` commands, respectively.

4.1. Modifying the Data to Use STATA

For our code examples, we assume that users have downloaded the example data file (`data_mi_2010_2012.dta`) and stored it in a local directory, which we refer to as `C:\mi`.

The STATA Multiple Imputation procedures (StataCorp, 2021) require that the original and the multiply-imputed data sets are included as one data set. We have created a data set that meets the requirements. The MI data set includes the original data set (before imputation) and the three STATA MI "system" variables, namely `_mi_m`, `_mi_id`, and `_mi_miss`.

- `_mi_m` is a data set indicator containing values $m = 0, 1, 2, ..., M$, where $m = 0$ represents the original data set, and $m = 1, 2, ..., M$ represent the imputed data sets.
- `_mi_id` is a unique ID variable.
- `_mi_miss` is the missing indicator variable (1 = missing, 0 = not missing).

Please note that in the original data set ($m = 0$) only the variables `_mi_m` and `_mi_id` contain values. The rest of the variables were set to missing (".") because the values belonging to those variables were not needed for the `mi estimate` procedures.

```
* Get the data.
global data_dir = "C:\mi"
use "$data_dir\data_mi_2010_2012.dta", clear
browse if Replicate == 1
```

	firm_size	hr_wage	pid	Replicate	year	_mi_m	_mi_id	_mi_miss
1501	10690	22.22	388	1	2010	1	961	.
1502	16068	29.29	115	1	2010	1	52	.
1503	6374	22.22	13	1	2010	1	109	.
1504	5527	12.12	348	1	2010	1	829	.
1505	2467	21.21	154	1	2010	1	184	.
1506	2552	9.09	319	1	2010	1	730	.
1507	11809	48.48	399	1	2010	1	997	.
1508	4712	24.24	343	1	2010	1	814	.
1509	19259	42.42	34	1	2010	1	808	.
1510	23209	26.26	197	1	2010	1	325	.
1511	11900	11.11	57	1	2010	1	1360	.

4.2. Estimating Means and Standard Errors

The code that follows shows how to compute means and standard errors for the variables `hr_wage` and `firm_size` for the year 2010 from the multiply-imputed data set. The printout below shows the computed means of both variables for 2010.

```
* Compute mean for the year 2010
  mi estimate: mean hr_wage firm_size if year == 2010
```

```
Multiple-imputation estimates     Imputations      =         10
Mean estimation                   Number of obs    =         500
                                  Average RVI      =      0.1750
                                  Largest FMI      =      0.2384
                                  Complete DF      =         499
DF adjustment:   Small sample     DF:     min      =      118.70
                                          avg      =      307.86
Within VCE type:     Analytic             max      =      497.01
```

	Mean	Std. Err.	[95% Conf.	Interval]
hr_wage	25.21768	.525444	24.18531	26.25005
firm_size	14568.08	522.3108	13533.83	15602.34

4.3. Estimating Model Coefficients

The code that follows shows how to run a regression model that includes the variable `hr_wage` as the outcome variable and `firm_size` as the predictor. Here, we estimate model coefficients for 2010.

```
    * Compute regression using the multiply-imputed data set
    mi estimate: regress hr_wage firm_size if year == 2010
```

```
Multiple-imputation estimates                 Imputations      =         10
Linear regression                             Number of obs    =         500
                                              Average RVI      =      0.0472
                                              Largest FMI      =      0.0398
                                              Complete DF      =         498
DF adjustment:   Small sample                 DF:     min      =      440.59
                                                      avg      =      452.36
                                                      max      =      464.13
Model F test:       Equal FMI                 F(   1,   440.6)   =       85.98
Within VCE type:          OLS                 Prob > F         =      0.0000
```

hr_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
firm_size	.0004475	.0000483	9.27	0.000	.0003527	.0005424
_cons	18.69884	.8537775	21.90	0.000	17.02109	20.37659

Conclusion

The CenHRS crosswalk is constructed using a mix of both deterministic and probabilistic matching of the employers of HRS survey respondents to Census business data. This type of data set contains multiply imputed data sets stacked in a single data file and requires MI procedures to produce valid statistical inferences. The SAS and STATA MI examples outlined in this document provide guidance on applying appropriate MI procedures to CenHRS data or other MI data sets.

References

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* John Wiley.

StataCorp. (2021). *Stata multiple-imputation reference manual: Release 17.* Stata Press. https://www.stata.com/manuals/mi.pdf