

Using SAS® for Multiple Imputation and Analysis of Longitudinal Data

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

ABSTRACT

“Using SAS for Multiple Imputation and Analysis of Data” presents use of SAS to address missing data issues and analysis of longitudinal data. Appropriate multiple imputation and analytic methods are evaluated and demonstrated through an analysis application using longitudinal survey data with missing data issues.

The analysis application demonstrates detailed data management steps required for imputation and analysis, multiple imputation of missing data values, subsequent analysis of imputed data, and finally, interpretation of longitudinal data analysis results. Key SAS tools including data step operations to produce needed data structures and use of PROC MI, PROC MIANALYZE, PROC MIXED, and PROC SGPLOT are highlighted.

INTRODUCTION

This paper presents a discussion of missing data issues, evaluation of multiple imputation and analysis methods appropriate for longitudinal data, and a detailed analysis application using SAS tools.

The application uses data from the Panel Study of Income Dynamics (PSID). This rich, longitudinal survey includes information on the family head's income for many years and presents an opportunity to employ descriptive methods and regression techniques, once missing data problems are addressed. SAS DATA STEP tools for data restructuring and use of PROC MI, PROC MIANALYZE, PROC SGPLOT and PROC MIXED are highlighted in the application.

The content of paper 1738 is useful to those working with longitudinal data with missing data problems and the need for an integrated process for multiple imputation and analysis of completed longitudinal data sets. Skills taught include how to use SAS to address missing data in longitudinal data sets, analyze completed data sets using appropriate descriptive and regression techniques, and combine results while accounting for both imputation variability and if needed, complex sample design features for correct variance estimation.

SECTION 1-OVERVIEW OF MISSING DATA AND MULTIPLE IMPUTATION

MISSING DATA IN LONGITUDINAL DATA SETS

Missing data is especially common in longitudinal data sets. Reasons for the missingness might be respondent attrition, survey structure where some questions are asked only of a sub-set of respondents, file-matching issues where some variables are not included in all data sets used in a merge process, and refusal to answer sensitive questions such as income, illegal behaviors, or certain health conditions.

Missing data is generally categorized as Missing at Random (MAR), Missing Completely at Random (MCAR), or Not Missing at Random (NMAR). See the SAS/STAT PROC MI documentation, Rubin (1987), Schafer (1997), or Raghunathan (2016) for more on these topics.

MULTIPLE IMPUTATION OF MISSING DATA

Multiple Imputation is a robust and flexible option for handling missing data. For longitudinal data as well as other 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).

Two important advantages of multiple imputation are: 1) MI incorporates the variability introduced by the imputation during variance estimation and 2) MI offers use of appropriate statistical models for generating plausible distributions of values to replace item-missing data. For more on multiple imputation and other types of imputation methods, see Schafer (1999), Rubin (1987) or more recently, Van Buuren (2012).

MULTIPLE IMPUTATION METHODS FOR LONGITUDINAL DATA

MI methods for longitudinal data can differ from those used to impute say, cross-sectional data. Typically, multiple records/values per respondent are included in longitudinal data sets (i.e. for each construct and time point) and correlation among repeated records should be captured by the MI process.

Longitudinal data is often collected in a long file format similar to Figure 1 where each respondent has one or more records representing a key construct over time. In this simple example, income for 2 years is collected in a multiple record per ID data format.

Long Format		
ID	Income	Year
1	100000	2013
1	200000	2015
2	150000	2013
2	200000	2015

Figure 1. Example of Long File

The long format is convenient for data collection and analysis but may not be appropriate for multiple imputation, thus data restructuring from long to wide or the reverse is often needed for multiple imputation and subsequent MI analyses. For more on MI of longitudinal data and model assumptions, see Raghunathan (2016) pages 121-126.

“Just Another Variable” Method

A popular MI method used with longitudinal data is called “Just Another Variable” or JAV, (Raghunathan, 2016). This method involves imputing missing data in a wide format data set with variables “strung out” one the data record. (See Figure 2). Though this method does not capture within-individual changes over time, it offers a convenient and flexible method for dealing with a varying number of records per individual and differing time points for follow-up data collection. In addition, it is easily performed in SAS

using PROC MI and PROC MIANALYZE. The JAV method is demonstrated in the analysis application in Section 2 of this paper.

As the name suggests, this method treats each variable as “just another” to be imputed. For example, the long data format in Figure 1 can be easily restructured into a wide format where multiple records are turned into one record per unique ID with each construct and time (income and year respectively) represented by uniquely named variables (Income_2013 and Income_2015).

Wide Format		
ID	Income_2013	Income_2015
1	100000	200000
2	150000	200000

Figure 2. Example of Long File Converted to Wide Format

This data structure permits multiple imputation of item-missing data for each respondent’s uniquely named variables in the rectangular data array. Once imputation is finished, the wide data set is generally “reversed” back to the long format for subsequent analysis of imputed longitudinal data.

Two-Fold Fully Conditional Specification Method

An alternative imputation method for longitudinal data is the two-fold fully conditional specification (FCS) approach proposed by Welch, Bartlett, Peterson, (2014). This method performs a two step or “two-fold” multiple imputation process outlined in Figure 3 (graphic adapted from Nevalainen, et al. (2009)).

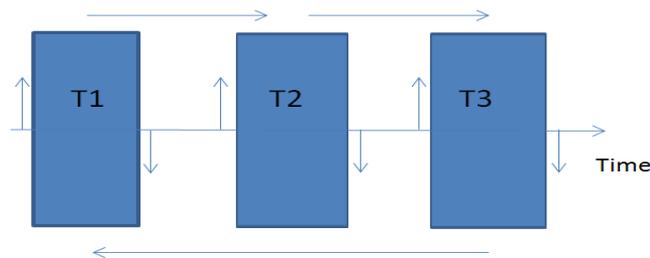


Figure 3. Diagram of Two-Fold Fully Conditional Specification Imputation

Figure 3 highlights how the two-fold FCS method first imputes missing data within each wave (up/down arrows around each box in figure below) and secondly, imputes across waves using a user-specified $t \pm k$ iterative process (horizontal arrows across top and bottom of figure). This method incorporates the impact of each individual’s responses at time t , and those around t , by using $t-k$ and $t+k$, where k is typically 1 or 2 (specified by analyst).

This method is not currently available in PROC MI but is discussed in detail in Nevalainen et al (2009). For more on the JAV and two-fold FCS methods, see De Silva, Moreno-Betancur, Madhu De Livera, Lee, and Simpson, (2017).

PLANNING FOR MULTIPLE IMPUTATION

Careful planning of a multiple imputation session is critical to produce both high quality imputations and subsequent analyses of imputed data sets. Often, the analyst is tempted to rush into multiple imputation without a complete understanding of the missing data problem and associated issues. The checklist presented in Table 1 is a suggested guide for planning the multiple imputation project.

Checklist of Issues and Considerations for the Multiple Imputation Process
Type of missing data - Item v. Unit, item missing data is topic of this presentation, unit generally handled by weighting adjustments
Assumptions – Missing at Random (MAR=default assumption of PROC MI/PROC MIANALYZE), Missing Completely at Random, Missing Not at Random
Types of variables imputed - Continuous, nominal, binary, ordinal, count/mixed
Missing data patterns - Arbitrary, monotone, file-matching
Amount of missing information - Extent of missing information is important factor when selecting M=(number of imputations)
Imputation model - Imputation model(s) depends on type of variable(s) that require imputation (continuous, categorical, count, etc.), should include all analysis variables plus additional “auxilliary” variables to enrich imputation models (include complex sample design variables and weights too, if applicable), each variable to be imputed may have a different imputation model
Number of imputations - Depends on how much data is missing, expected relative efficiency, how many records and variables included in imputation models, and other factors, may be an iterative process to evaluate M at certain numbers (say 10, 20, 25, 50...), if in doubt, use a higher rather than lower M
Big Data Imputation - Hardware/software limits, how many variables/records realistic to impute/analyze, end-user capacities and analytic usage, burden on imputer and analyst

Table 1. Checklist of Issues and Considerations for the Multiple Imputation Process

SECTION 2 - ANALYSIS APPLICATION

MULTIPLE IMPUTATION AND ANALYSIS OF LONGITUDINAL SURVEY DATA FROM THE PANEL STUDY OF INCOME DYNAMICS (PSID)

Data Management

The analysis application uses data from the Panel Study of Income Dynamics (PSID), <https://psidonline.isr.umich.edu/>. The PSID is a long-running (1968-present) longitudinal study of U.S. families. The source data for this example was downloaded from the PSID data center.

The analysis application employs descriptive statistics and linear growth curve models to analyze the family head’s wages/salary over time (1997-2013, odd years) by completed college status (completed grade 16+ US education system) while using multiply imputed data in all analyses.

The downloaded PSID data set was filtered to include individuals that meet all three conditions, n=2.267:

- Individuals must be the family head in each odd year from 1997-2013 (9 waves of data),
- From either the Survey Research Center (SRC) or U.S. Census (Census) samples from 1968,
- Head must be present in family in each year of series.

Prior to multiple imputation of item-missing data and subsequent analysis of completed data sets, some data management was needed. The data from the PSID data center was structured as a wide file, therefore no transposition was required for imputation of missing data. However, previously imputed values (done by PSID staff using a modified “hotdeck” imputation method) were returned to their original missing data values with the goal of using the preferred multiple imputation method rather than the hotdeck method.

A few additional variables were created for use in the multiple imputation process. For example, the natural log of head’s wages/salary was created to address non-normal distributions, a combined Stratum and Sampling Error Computing Unit (SECU) or PSU variable was constructed to incorporate the impact of the PSID complex sample design stratification/clustering features in the imputation models, and a series of imputation “flag” variables were created to identify imputed values and assist in imputation diagnostics. For more on inclusion of complex sample design variables in multiple imputation, see Berglund and Heeringa (2014) or Reiter, Raghunathan, and Kinney (2006). Note that the SAS code for preliminary data management is not shown here but available from the author upon request.

The filtered, analysis data set includes the variables listed in Table 2, n=2,267. Note that age, education (highest grade completed), and log head’s wages/salary include a uniquely named variable for each of nine waves of data used in the analysis application. The other variables are time-invariant and do not require strings of uniquely named variables. Also, variables highlighted in red have some missing data.

Contents of Final MI Data Set (Wide Format)
Er32000 - Gender (1=M, 2=F), fully observed
Age1-Age9 - Age in 1997, 1999, 2001, 2003, 2005, 2007, 2009, 2011, 2013, fully observed
Strat_psu – Combined stratum and SECU (PSU) variable, fully observed, used to incorporate complex sample design features in imputation model
Er34268 – Probability weight from 2013, fully observed
Ed1-Ed9 – Highest grade completed (odd years 1997- 2013), missing data on each variable
Loghdwg1-Loghdwg9 – Log of head’s wages/salary (odd years 1997-2013), missing data on each variable
ID – ID68 and Person number combined to create unique individual identifier, fully observed
Samplecat – Sample indicator of SRC or Census (1968 original sample), fully observed data

Table 2. Contents of Final Analysis Data Set

MULTIPLE IMPUTATION AND ANALYSIS PROCESS

Prior to a detailed presentation of analysis examples, a brief outline of the MI process is provided.

Step 1 Multiple Imputation of Missing Data

Step 1 includes evaluation of the missing data problem, multiple imputation of missing data using PROC MI with the Fully Conditional Specification (FCS) method and other options/statements, and evaluation of imputations using diagnostic tools such as plots and comparisons of observed versus imputed values. Once the imputations are considered final, the imputed data set is converted from a wide to long format suitable for longitudinal MI data analysis.

Step 2 Analysis of Completed Data Sets

Step 2 consists of analysis of the long, imputed data set using either standard (based SRS samples) or SURVEY procedures designed for analysis of complex sample data. The examples presented in this

paper use standard procedures despite the complex sample design of the PSID. This is done to keep analysis examples relatively simple, however a repeat of Example 2 using SAS-callable IVEware with SASMOD is presented in Appendix B.

IVEware is a tool that can perform multiple imputation and analysis of imputed data while also handling complex sample design variance estimation. The software and documentation is freely available from iveware.org. The SASMOD command uses the correct combining rules for multiply imputed data and also implements Jackknife Repeated Replication (Rust, 1985) for design-based variance estimation. Selected SAS procedures such as PROC MIXED, PROC GENMOD and others can be used with SASMOD framework. This additional complexity is needed only for analysis of data derived from complex samples and also for SAS procedures such as PROC MIXED that do not have an equivalent SAS SURVEY procedure.

Step 3. Combine Results

PROC MIANALYZE combines the results of MI Steps 1 and 2 and generates valid statistical inferences by accounting for the variability introduced by the MI process.

THREE- STEP MULTIPLE IMPUTATION PROCESS

Step 1- Multiple Imputation of Missing Data

Evaluation of Missing Data Problem

Prior to imputation, evaluation of the extent of the missing data problem, types of variables to be imputed or used in imputation models, and the pattern of missing data is recommended.

Two approaches are demonstrated in the analysis examples: use of 1) PROC MEANS to examine each variable's observed n (n), number missing (nmiss), and mean, minimum, and maximum values and 2) PROC MI with NIMPUTE=0 to obtain a missing data patterns grid without imputation. Both approaches assist with evaluation of the missing data problem prior to multiple imputation:

```
proc means data=w.psid1 n nmiss mean min max;
  var er32000 age1-age9 strat_psu er34268 ed1-ed9 loghdwg1-loghdwg9;
run;

proc mi data=w.psid1 nimpute=0;
  var er32000 age1-age9 strat_psu er34268 ed1-ed9 loghdwg1-loghdwg9;
run;
```

PROC MEANS output is presented in Figure 4 and indicates missing data on each of the 9 education variables (ED1-ED9) and 9 log head's wages/salary variables (LOGHDWG1-LOGHDWG9). All other variables are fully observed.

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Minimum	Maximum
ER32000	SEX OF INDIVIDUAL	2267	0	1.2818703	1.0000000	2.0000000
age1	Age 1997	2267	0	43.2831936	18.0000000	84.0000000
age2	Age 1999	2267	0	45.3361270	20.0000000	86.0000000
age3	Age 2001	2267	0	47.4159682	22.0000000	88.0000000
age4	Age 2003	2267	0	49.4194971	24.0000000	90.0000000
age5	Age 2005	2267	0	51.3855315	26.0000000	92.0000000
age6	Age 2007	2267	0	53.4243494	28.0000000	94.0000000
age7	Age 2009	2267	0	55.4159682	30.0000000	97.0000000
age8	Age 2011	2267	0	57.4322894	32.0000000	98.0000000
age9	Age 2013	2267	0	59.4243494	34.0000000	100.0000000
strat_psu	Combined Strata and SECU	2267	0	1505.47	101.0000000	3202.00
ER34268	CORE/IMM INDIVIDUAL LONGITUDINAL WT 13	2267	0	43.3975827	0.7180000	196.4410000
ed1	Yrs Completed Ed 1997	2228	39	13.2163375	3.0000000	17.0000000
ed2	Yrs Completed Ed 1999	2228	39	13.2158887	3.0000000	17.0000000
ed3	Yrs Completed Ed 2001	2227	40	13.2177818	3.0000000	17.0000000
ed4	Yrs Completed Ed 2003	2227	40	13.2218231	3.0000000	17.0000000
ed5	Yrs Completed Ed 2005	2227	40	13.2236192	3.0000000	17.0000000
ed6	Yrs Completed Ed 2007	2227	40	13.2245173	3.0000000	17.0000000
ed7	Yrs Completed Ed 2009	2249	18	13.5028902	2.0000000	17.0000000
ed8	Yrs Completed Ed 2011	2248	19	13.5031139	2.0000000	17.0000000
ed9	Yrs Completed Ed 2013	2250	17	13.5342222	2.0000000	17.0000000
loghdwg1	Log Head Wages 1996	2182	85	8.0864335	0	13.2638959
loghdwg2	Log Head Wages 1998	2182	85	8.1917886	0	13.0815435
loghdwg3	Log Head Wages 2000	2183	84	8.1495913	0	13.6529928
loghdwg4	Log Head Wages 2002	2162	105	7.6936513	0	15.0682738
loghdwg5	Log Head Wages 2004	2166	101	7.8731075	0	14.7277937
loghdwg6	Log Head Wages 2006	2170	97	7.5520295	0	13.6876783
loghdwg7	Log Head Wages 2008	2163	104	7.3336556	0	14.0778756
loghdwg8	Log Head Wages 2010	2180	87	6.5490353	0	13.9639314
loghdwg9	Log Head Wages 2012	2190	77	6.2370929	0	15.0094333

Figure 4. Results from PROC MEANS

The education variables (ED1-ED9) represent highest grade completed and are treated as continuous. The continuous log head's wages/salary variables (LOGHDWG1-LOGHDWG9) represent the previous year's wages/salary and range from 0 (did not receive wage/salary in dollars for a given year) to 15.06, on the log scale.

We acknowledge that log-transformed variables can produce bias and heavy tails in the distribution of the back-transformed, imputed version but we proceed with the transformation as recent research has demonstrated that for regression estimates, this bias is often mild, von Hippel (2013). Another important caution is that age and time are linked and if age is used as a predictor in growth models, it should be treated as time-invariant, e.g., age at a fixed point such as age in 1997.

Figure 5 presents part of the Missing Data Patterns Grid produced by PROC MI. The grid provides frequency counts/percentages for observed data ("X") and missing data ("."), for each variable to be used in multiple imputation models.

Group	ER32000	age1	age2	age3	age4	age5	age6	age7	age8	age9	strat_psu	ER34268	ed1	ed2	ed3	ed4	ed5	ed6	ed7	ed8	ed9	loghdwg1	loghdwg2	loghdwg3	loghdwg4	loghdwg5	loghdwg6	loghdwg7	loghdwg8	loghdwg9	Freq	Percent		
1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	1815	80.06	
2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	23	1.01
3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	25	1.10	
4	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	4	0.18	

Figure 5. Results from PROC MI

There are 128 unique groups of missing/observed data combinations with just the first four presented here. Group 1 is defined as fully observed on all variables, i.e. 80.06% of sample or 1815 individuals are assigned to the complete data group while the entire grid identifies an additional 127 missing data patterns each with $\leq 1.5\%$ missing data, with an arbitrary missing data pattern.

Based on Figures 4 and 5, the missing data problem is understood. For example, there are 18 continuous variables that require imputation (constituting 20% of full data set), the fully observed variables are a mix of continuous and categorical variables, and the missing data pattern is arbitrary. This information helps determine proper settings in the PROC MI code for multiple imputation.

Multiple Imputation using PROC MI

Next, multiple imputation of missing data is performed by PROC MI with a number of statements and options. In the code below, the SEED=2017 option ensures the ability to replicate results at a later time, NIMPUTE=10 requests SAS create M=10, OUT=IMPUTE_PSID_MI saves a vertically stacked, imputed data set for subsequent use, ROUND=() formats imputed values for each corresponding variable in the VAR statement, the BY SAMPLECAT statement imputes missing values separately within the PSID SRC and Census samples, a CLASS statement declares ER32000 and STRAT_PSU as classification variables, FCS REGPMM selects the FCS Predicted Mean Matching method to impute missing data with NBITER=20 burn-in iterations and K=8 nearest neighbors, the (PLOTS=TRACE) option requests diagnostic trace plots for LOGHDWG1-LOGHDWG9, and the VAR statement lists variables to be included in the imputation models, ordered from no missing data to most missing, moving from left to right:

```
proc mi data=w.psid1
  seed=2017
  nimpute=10
  out=impute_psid_mi
  round= . . . . . 1 1 1 1 1 1 1 1 1 .01 .01 .01 .01 .01 .01
  .01 .01 .01 ;
  by samplecat ;
  class er32000 strat_psu ;
  fcs nbiter=20 regpmm(ed1-ed9 / k=8 ) ;
  fcs nbiter=20 plots=trace regpmm(loghdwg1-loghdwg9 / k=8 ) ;
  var er32000 age1-age9 strat_psu er34268 ed1-ed9 loghdwg1-loghdwg9 ;
run;
```

Multiple Imputation Diagnostics

Figures 6 and 7 present Trace plots of the 2004 head's wages/salary imputation. Because the imputations were done separately for the SRC and Census samples, we evaluate each plot separately. Trace plots are typically used to detect non-random patterns that may arise during the imputation process. Random variation around imputed values, once burn-in iterations are complete, generally indicates a lack of problems with the imputations.

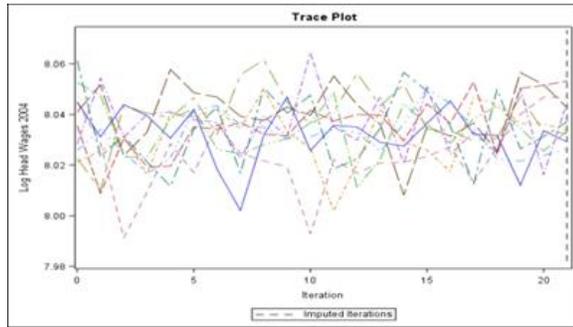


Figure 6. Trace Plot of Head's Wages/Salary 2004, SRC Sample

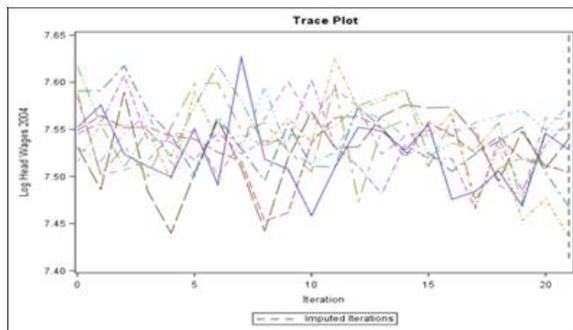


Figure 7. Trace Plot of Head's Wages/Salary 2004, Census Sample

Based on the random patterns of Figures 6 and 7, there is little concern about the quality of these imputations. Detailed evaluation of all 18 plots produced by PROC MI (not shown here) also indicates no issues with the imputations.

Alternatively, evaluation of imputations can be carried out by checking mean head's wages/salary by SAMPLECAT, _IMPUTATION_ (the automatic variable indicating imputation number), and the user-defined imputation flag, IMPHDWG1 (created in a previous DATA STEP):

```
proc means data=impute_psid_mi ;
  class samplecat _imputation_ imphdwg1 ;
  var loghdwg1 ;
run ;
```

Figure 8 presents mean 1996 head's wages/salary for 5 of 10 imputations by Census and SRC samples, imputation number, and imputation status. This diagnostic check also reveals no apparent issues between observed (imphdwg1=0) versus imputed (imphdwg1=1) values.


```

wgt=weight[i];
sex=er32000;
stratum=er31996;
cluster=er31997;
age = ag[i];
year = y[i];
time = t[i] ;
completeded=ed[i] ;
if ed[i] >= 16 then cg [i]=1 ; else cg [i]=0 ;
collegegrad=cg[i] ;
_mult_=_imputation_ ;
output ;
end ;
keep id hdwg headwage wgt sex stratum cluster age year time completeded
collegegrad _mult_ _imputation_ er30001 er30002 er32000 samplecat er34268
;
run ;

```

With the imputed, long data set now ready for MI analysis, detailed SAS code and results are presented for typical longitudinal analysis techniques.

REVIEW OF ANALYSIS EXAMPLES

The primary analytic goal is to examine trends in head's wages/salary over time by college graduation status. Both descriptive and regression techniques are used to address this goal.

The descriptive analysis focuses on mean head's wages/salary by year and college graduation status, produced by PROC MEANS and PROC GPLOT.

Growth models are executed using PROC MIXED with a RANDOM statement to account for within and between-subject variation where predicted head's wages/salary (based on mixed model results) are calculated in the DATA STEP and then used in plots produced by PROC SGPLOT.

As previously explained, Examples 1 and 2 ignore the PSID complex sample design features but incorporate all appropriate MI analysis and combining techniques. However, to provide guidance for those working with complex sample data and multiply imputed data, Example 2 is repeated using IVEware and the SASMOD command with PROC MIXED to compare how variances change when the complex sample design features of the PSID are incorporated. Results are presented in Appendix B.

Analysis Example 1 - Mean Head's Wages/Salary by Year and College Graduation Status

Step 2. Analysis of Completed Data Sets

Example 1 inputs the long, imputed data set generated in MI Step 1 and demonstrates descriptive analysis of head's wages/salary by imputation, college graduation status, and year. PROC MEANS is used to prepare summary statistics that are saved to an output data set for use in PROC MIANALYZE.

The following code first sorts the LONG_IMPUTED data set by _IMPUTATION_, COLLEGEGRAD, and TIME. Next, PROC MEANS with BY and WEIGHT statements is used to obtain weighted means of head's wages/salary within each of 10 imputed data sets, by college status and time. The OUTPUT statement saves the statistics of interest to a file called AVGWAGE and finally, PROC SORT sorts the data for use in PROC MIANALYZE:

```

proc sort data=w.long_imputed ;
  by _imputation_ collegegrad time ;
run ;

```

```

* Run PROC MEANS in long data set with 10*2267 (10 imputations) * 9 records
per individual = 204,030 ;
proc means data=w.long_imputed mean stderr ;
  by _imputation_ collegegrad time ;
  var headwage ;
  weight er34268 ;
  output out=avgwage mean=mean_headwage stderr=se_headwage ;
run ;

proc print data=avgwage ;
run ;

proc sort data=avgwage ;
  by collegegrad time _imputation_ ;
run ;

```

Step 3 – Combine Results

PROC MIANALYZE combines results from Step 2 and generates variances that account for the additional variability introduced by the multiple imputation. These combined estimates are then used to plot mean wages/salary over time by college graduation status.

The following code invokes PROC MIANALYZE and includes a BY statement to produce a combined estimates by college status and time, MEAN_HEADWAGE is declared as the MODELEFFECTS variable, SE_HEADWAGE as the STDERR variable, and ODS OUTPUT saves an output data set with combined “parameter estimates” or means in this case, for use in PROC SGPLOT.

PROC FORMAT creates descriptive labels for the plot, while PROC SGPLOT inputs the data set from PROC MIANALYZE and uses SERIES, XAXIS, YAXIS, and FORMAT statements to customize the plot presented in Figure 9:

```

proc mianalyze data=avgwage ;
  by collegegrad time ;
  modeleffects mean_headwage ;
  stderr se_headwage ;
  ods output parameterestimates=outcombine_1 ;
run ;

proc format ;
  value tf 1='1997' 2='1999' 3='2001' 4='2003' 5='2005' 6='2007' 7='2009'
8='2011' 9='2013' ;
  value cf 0='No' 1='Yes' ;
run ;

proc sgplot data=outcombine_1 ;
title "Mean Head's Wages/Salary by College Graduate Status" ;
  series x=time y=estimate / group=collegegrad markers ;
  xaxis label='Year' ;
  yaxis label='Mean Head Wage/Salary 1997 to 2013 (in 2013 Dollars)' ;
  format time tf. collegegrad cf. ;
run ;

```

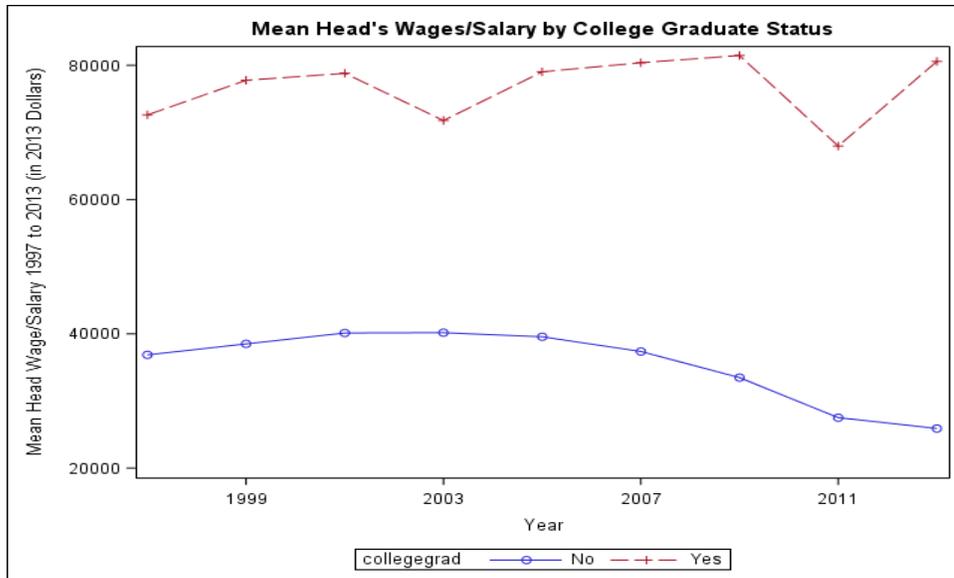


Figure 9. Mean Head's Wages/Salary by College Graduate Status, 1997-2013

Figure 9 plots head's wages/salary trends over time for college graduates and non-graduates and suggests a possible interaction between year and college graduate status. For college graduates, there is a mix of positive and sharp negative slopes while for non-graduates, the slopes are flatter/smaller and primarily negative. The plot highlights trends as heads aged and experienced a changing economic climate and other changes in their economic circumstances between 1997-2013. Furthermore, a wage differential of about \$33,000 between college graduates and those that did not graduate from college persists over the years of interest.

Analysis Example 2 – Growth Model

Step 2. Analysis of Completed Data Sets

Example 2 demonstrates use of a growth model to investigate the impact of time and college graduation status on head's wages/salary. This model accounts for between-subject (intercept) and within-subject (time) variation by requesting random intercepts and slopes while using time as a continuous rather than categorical predictor, as in Example 1. As in the previous example, the long, imputed data set from Step 1 is used as input for this example.

Model Fitting Prior to Inference

Prior to the inference step or Step 4 of the model building process, model fitting was performed by following Steps 1-3 as recommended by the SAS Institute "Mixed Model Analyses of Repeated Measures Data" course notes. (Results not shown but available upon request).

- Step 1- Model mean structure, specify fixed effects
- Step 2- Set covariance structure for within-subject and/or between-subject effects
- Step 3- Use Generalized Least Squares (GLS) to fit mean model with selected covariance structure
- **Step 4- Make statistical inference based on Step 3, aim for parsimonious model**

To summarize, model fitting was done separately within each of the M=10 imputed data sets with 3 covariance structures: 1) UN, 2) AR(1), and 3) Toeplitz. PROC MIANALYZE was used to combine results for 10 imputed data sets/models and incorporate MI variability in variance estimates.

After evaluation of the three covariance structures using AIC and BIC statistics, unstructured (UN) covariance was selected as the preferred structure for inference.

Step 4 - Inference

The following code uses PROC MIXED with a number of options and statements. Use of BY `_IMPUTATION_` executes a growth model within each of 10 imputed data sets, the CLASS statement treats college graduation status and the respondent ID variable as classification variables, the MODEL statement uses HEADWAGE as a continuous outcome with continuous TIME and the interaction of TIME*COLLEGEGRAD with a SOLUTION option to request fixed effects solutions and DDFM=BW to request the between-within method for computing denominator degrees of freedom, and a RANDOM INTERCEPT TIME / TYPE=UN SUBJECT=ID statement to request random intercepts and slopes with an unstructured covariance matrix. The PSID 2013 longitudinal weight is used in the WEIGHT statement while ODS OUTPUT requests an output data set of parameter estimates for use in PROC MIANALYZE. The final section of code uses PROC PRINT to show the contents of the output data set OUTCOMBINE_RANDOM:

```
*Step 4 Inference: Use RANDOM INTERCEPT/SLOPE with unstructured covariance;
proc mixed data=w.long_imputed noclprint;
  by _imputation_;
  class collegegrad id;
  model headwage = time collegegrad time*collegegrad / solution ddfm=bw;
  random intercept time / type=un subject=id;
  weight er34268;
  ods output solutionf=outcombine_random;
run;

proc print data=outcombine_random;
run;
```

<code>_Imputation_</code>	Effect	collegegrad	Estimate	StdErr	DF	tValue	Probt
1	Intercept	_	71300	2530.22	2266	28.18	<.0001
1	time	_	-1649.66	443.64	18E3	-3.72	0.0002
1	collegegrad	0	-29708	3009.02	156	-9.87	<.0001
1	collegegrad	1	0
1	time*collegegrad	0	913.39	524.87	18E3	1.74	0.0818
1	time*collegegrad	1	0
2	Intercept	_	71764	2558.86	2266	28.05	<.0001
2	time	_	-1696.62	447.75	18E3	-3.79	0.0002

Imputation	Effect	collegegrad	Estimate	StdErr	DF	tValue	Probt
2	collegegrad	0	-29651	3044.51	153	-9.74	<.0001
2	collegegrad	1	0
2	time*collegegrad	0	873.99	530.78	18E3	1.65	0.0997
2	time*collegegrad	1	0

Table 3. Print-Out of Fixed Effects Parameters for 2 of 10 Imputed Data Sets

Table 3 displays partial output for 2 of 10 imputations and includes fixed effects regression estimates, standard errors, degrees of freedom, t values, and p values. Note that the estimates and associated statistics are slightly different for each imputed data set, reflecting the differing imputed values. This data set serves as input for PROC MIANALYZE during the combining step.

Step 3 – Combine Results

The next block of code details use of PROC MIANALYZE to combine results from MI Step 2. The DATA=OUTCOMBINE_RANDOM statement inputs the data set saved in Step 2, the PARS(MCLASSVAR=FULL) statement declares a full set of discrete levels for the CLASS variables, a CLASS statement declares COLLEGEGRAD as a categorical variable, MODELEFFECTS declares the intercept and other predictor variables as model effects, and ODS OUTPUT creates a data set of estimates called OUTCOMBINE_RANDOM_A. PROC PRINT produces a listing of the contents of the final output data set:

```
proc mianalyze parms(classvar=full)=outcombine_random;
  class collegegrad;
  modeleffects intercept time collegegrad time*collegegrad;
  ods output parameterestimates=outcombine_random_a;
run;
proc print noobs data=outcombine_random_a;
  var parm collegegrad estimate stderr tvalue probt;
run;
```

Parameter	College Graduate	Estimate	MI SE	T Value	P value
intercept	.	71442.00	2553.67	27.98	<.0001
time	.	-1678.08	447.16	-3.75	0.0002
collegegrad	0	-29798.00	3027.97	-9.84	<.0001
collegegrad	1.000000	0	.	.	.
time*collegegrad	0	924.72	529.55	1.75	0.0808
time*collegegrad	1.000000	0	.	.	.

Table 4. Combined Parameter Estimates for Growth Model

Table 4 presents combined parameter estimates, standard errors, and *t* and *p* values. These results are correctly combined and account for the variability introduced by the MI process.

In addition, model estimates account for between-subject (intercept) and within-subject (time) variation through use of the RANDOM statement in PROC MIXED. The results from Table 4 suggest that time, college graduation status and their interaction are all significant at the alpha=0.10 level while time and college status are also significant at the alpha=0.05 level. Because the interaction term is nearly significant at the 0.05 level, it remains in the model for demonstration purposes.

Plot Predicted Head's Wages/Salary by College Graduate Status and Year

Based on the combined parameter estimates above, predicted head's wages/salary is calculated manually in the DATA STEP and subsequently used in PROC SGLOT to produce Figure 10, a plot of regression slopes for predicted wages/salary by college graduation status:

```
data predicted_random ;
  set w.long_imputed ;
  if collegegrad=1 then predicted_hdwage_random=71442 + time * -1678;
  else if collegegrad=0 then predicted_hdwage_random=(71442-29798) +
time*(-1678 + 925);
run ;

proc format ;
  value tf 1='1997' 2='1999' 3='2001' 4='2003' 5='2005' 6='2007' 7='2009'
8='2011' 9='2013' ;
  value cf 0='No' 1='Yes' ;
run ;

title "Predicted Head's Wages by College Graduate Status and Time, Model
includes Random Statement and Results Combined by PROC MIANALYZE" ;
proc sgplot data=predicted_random ;
  series x=time y=predicted_hdwage_random / group=collegegrad ;
  xaxis type=discrete label='Time';
  yaxis label='Predicted Mean Head Wage/Salary 1997 to 2013 (in 2013
Dollars)' ;
  format collegegrad cf. time tf. ;
run ;
```

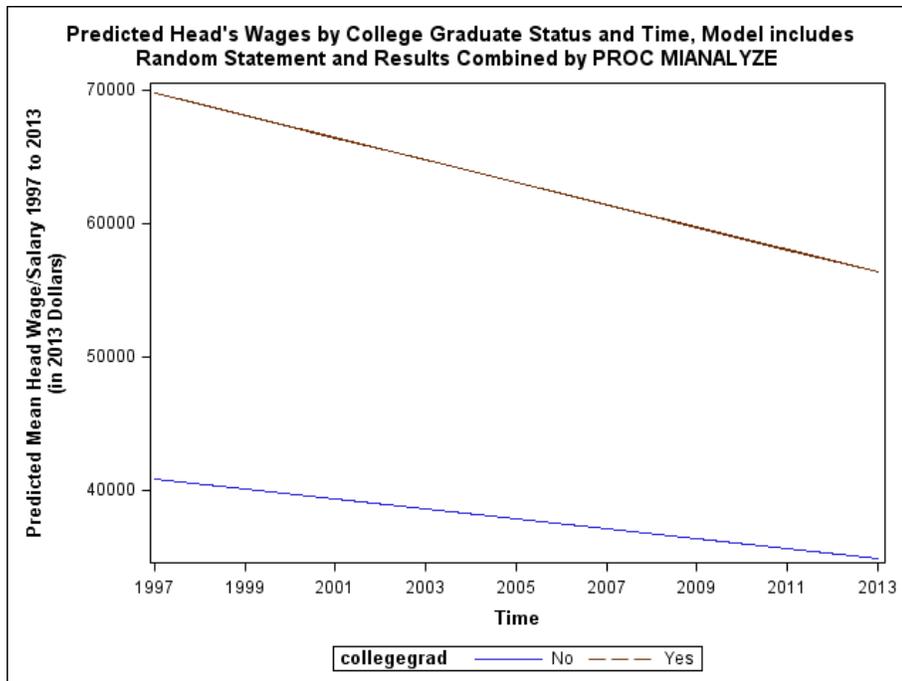


Figure 10. Predicted Head's Wages/Salary by College Graduate Status and Year

Figure 10 presents a plot of regression lines of predicted head's wages/salary (1997-2013) for college graduates and non-college graduates. The negative slope (head's wages/salary are in 2013 dollars) for college graduates is much steeper than for non-graduates but the intercepts are estimated to be about \$30,000 lower for non-graduates. Though non-graduates have a flatter slope, their income over time is much lower than college graduates, reflecting wage/salary differentials between levels of education over 1997-2013, among US household heads.

Again, these results are derived from an analysis that accounts for the MI process in variance estimates but does not incorporate the complex sample design features. For a repeat of this example using IVEware with the SASMOD command, see Appendix B.

CONCLUSION

This presentation has covered missing data issues in longitudinal data, considered two potential imputation methods appropriate for longitudinal data, demonstrated multiple imputation using one of the proposed methods, and provided a detailed analysis application.

The application uses complex sample survey data from PSID to study wages/salary trends as US household heads age over the years 1997-2013. The application demonstrates the three step process: 1) use PROC MI to perform multiple imputation, 2) analysis of completed data sets using growth models (PROC MIXED/PROC SGPLOT) and descriptive techniques (PROC MEANS/PROC SGPLOT), and 3) combine analyses of imputed data sets using PROC MIANALYZE. The growth model example is also repeated using the IVEware/SASMOD command to demonstrate MI combining and Jackknife Repeated Replication variance estimation for complex sample survey data.

REFERENCES

- Berglund and Heeringa, “Multiple Imputation of Missing Data Using SAS” (2014)
- Heeringa, West, and Berglund, “Applied Survey Data Analysis” 2nd Edition, Chapman Hall, (2017)
- Hill, “The Panel Study of Income Dynamics” (1991)
- IVEware Software and Support: <http://www.src.isr.umich.edu/software/>
- Johnson, “Paper P-702 Individual Growth Analysis Using PROC MIXED”, (2001), analytics.ncsu.edu/sesug/2001/P-702.pdf
- Nevalainen, Kenward, Vertanin, “Missing values in longitudinal dietary data: a multiple imputation approach based on a fully conditional specification” (2009)
- Panel Study for Income Dynamics (PSID): <https://psidonline.isr.umich.edu/>
- Raghunathan, “Missing Data Analysis in Practice” (2016)
- Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology (2001)
- Rubin, D., Multiple Imputation for Nonresponse in Surveys (1987)
- Rust, K. (1985), “Variance Estimation for Complex Estimators in Sample Surveys,” Journal of Official Statistics, 1, 381–397.
- SAS Institute Training Course Notes for Mixed Model Analysis
- SAS Support: <https://support.sas.com/en/support-home.html>
- Anurika Priyanjali De Silva, Margarita Moreno-Betancur, Alysha Madhu De Livera, Katherine Jane Lee and Julie Anne Simpson, “A comparison of multiple imputation methods for handling missing values in longitudinal data in the presence of a time-varying covariate with a non-linear association with time: a simulation study” (2017)
- von Hippel, “Should a Normal Imputation Model Be Modified to Impute Skewed Variables” (2013)

ACKNOWLEDGMENTS

Thanks to Dr. Steven Heeringa, Dr. Martha Hill, both from UM-ISR and Rob Agnelli of SAS Technical Support for their generous support.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Patricia Berglund
University of Michigan-Institute for Social Research
pberg@umich.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

APPENDIX A – SAS CODE

```
*P. Berglund 2018 SGF SAS Code for Paper 1738 ;
libname w 'P:\SAS Global Forum 2018' ;

* Examine missing data problem ;
proc means data=w.psid1 n nmiss mean min max ;
  var er32000 age1-age9 strat_psu er34268 ed1-ed9 loghdwg1-loghdwg9 ;
run ;

title "Missing Data Grid from PROC MI";
proc mi data=w.psid1 nimpute=0 ;
  var er32000 age1-age9 strat_psu er34268 ed1-ed9 loghdwg1-loghdwg9 ;
run ;

* Use PROC MI with FCS and trace plots for imputation, with BY statement to
do separate imputations within SRC and CENSUS samples ;
proc sort data=w.psid1 ;
  by samplecat ;
run ;

title "Impute by Sample Category (SRC and Census Separately), FCS Method
with PMM with Round=.01, K=8, M=10 and Seed=2017 " ;
ods graphics on ;
proc mi data=w.psid1
  seed=2017
  nimpute=10
  out=impute_psid_mi
  round= . . . . . 1 1 1 1 1 1 1 1 1 .01 .01 .01 .01 .01
.01 .01 .01 .01 ;
  by samplecat ;
  class er32000 strat_psu ;
  fcs nbiter=20 regpmm(ed1-ed9 / k=8 ) ;
  fcs nbiter=20 plots=trace regpmm(loghdwg1-loghdwg9 / k=8 ) ;
  var er32000 age1-age9 strat_psu er34268 ed1-ed9 loghdwg1-loghdwg9 ;
run;

proc means data=impute_psid_mi ;
  class samplecat _imputation_ imphdwg1 ;
  var loghdwg1 ;
run ;

* Create a long data set with multiple records per person within each
imputed data set (_imputation_), using completed data sets ;
data long_imputed ;
  set impute_psid_mi ;
  if samplecat='SRC' then src=1 ; else src=0 ;
  if er32000=2 then female=1 ; else female=0 ;
  id=er30001*10000 + er30002 ;

* use arrays to output multiple records per individual ;
array w [*] loghdwg1-loghdwg9 ; * head wages for each year ;
array ed [*] ed1-ed9 ; * completed education for each year ;
array cg [*] cg1-cg9 ; * 16+ years of completed education ;
```

```

array y [9] _temporary_ (1997 1999 2001 2003 2005 2007 2009 2011 2013) ;
* create year of study ;
array wi [9] _temporary_ (1.45 1.40 1.32 1.27 1.19 1.12 1.09 1.04 1.00) ;
* convert each yr to 2013 dollars ;
array t [9] _temporary_ (1 2 3 4 5 6 7 8 9) ; * time is set from 1 to 9 ;
array weight [*] er33430 er33546 er33637 er33740 er33848 er33950 er34045
er34154 er34268 ; * longitudinal individual weight by yr ;
array ag [*] age1-age9 ; * age in each yr ;

do i = 1 to 9 ;
hdwg=exp (w[i]) ;
headwage=hdwg * wi[i] ;
wgt=weight[i];
sex=er32000;
stratum=er31996;
cluster=er31997;
age = ag[i];
year = y[i];
time = t[i] ;
completeded=ed[i] ;
if ed[i] >= 16 then cg [i]=1 ; else cg [i]=0 ;
collegegrad=cg[i] ;
_mult_=_imputation_ ;
output ;
end ;
keep id hdwg headwage wgt sex stratum cluster age year time completeded
collegegrad _mult_ _imputation_ er30001 er30002 er32000 samplecat er34268
;
run ;

* Example 1 - Plot of mean head's wages over time, descriptive analysis
using imputed data sets and combining with PROC MIANALYZE ;
proc sort data=w.long_imputed ;
by _imputation_ collegegrad time ;
run ;

ods html style=sasweb;
* Run PROC MEANS in long data set with 10*2267 (10 imputations) * 9 records
per individual = 204,030 ;
proc means data=w.long_imputed mean stderr ;
by _imputation_ collegegrad time ;
var headwage ;
weight er34268 ;
output out=avgwage mean=mean_headwage stderr=se_headwage ;
run ;
proc print data=avgwage ;
run ;

proc sort data=avgwage ;
by collegegrad time _imputation_ ;
run ;

* PROC MIANALYZE to Combine Results from PROC MEANS ;
proc mianalyze data=avgwage ;
by collegegrad time ;
modeleffects mean_headwage ;

```

```

stderr se_headwage ;
ods output parameterestimates=outcombine_1 ;
run ;

* Plot Means by Sample and Time 1997-2013 1=1997, 2=1999 etc. ;
proc format ;
value tf 1='1997' 2='1999' 3='2001' 4='2003' 5='2005' 6='2007' 7='2009'
8='2011' 9='2013' ;
value cf 0='No' 1='Yes' ;
run ;

proc sgplot data=outcombine_1 ;
title "Mean Head's Wages/Salary by College Graduate Status" ;
series x=time y=estimate / group=collegegrad markers ;
xaxis label='Year' ;
yaxis label='Mean Head Wage/Salary 1997 to 2013 (in 2013 Dollars)' ;
format time tf. collegegrad cf. ;
run ;

* Example 2-Step 4 Inference: Use RANDOM INTERCEPT / SLOPE with
unstructured covariance ;
proc mixed data=w.long_imputed noclprint ;
by _imputation_ ;
class collegegrad id ;
model headwage = time collegegrad time*collegegrad / solution ddfm=bw ;
random intercept time / type=un subject=id ;
weight er34268 ;
ods output solutionf=outcombine_random ;
run ;

proc print data=outcombine_random ;
run ;

* PROC MIANALYZE to Combine Results from PROC MIXED with random intercept
and slope;
proc mianalyze parms(classvar=full)=outcombine_random ;
class collegegrad ;
modeleffects intercept time collegegrad time*collegegrad ;
ods output parameterestimates=outcombine_random_a ;
run ;

proc print data=outcombine_random_a noobs ;
var parm collegegrad estimate stderr tvalue probt ;
run ;

data predicted_random ;
set w.long_imputed ;
if collegegrad=1 then predicted_hd wage_random=71442 + time * -1678 ;
else if collegegrad=0 then predicted_hd wage_random = (71442-29798) +
time*(-1678 + 925) ;
run ;

* Plot Means by Sample and Time 1997-2013 1=1997, 2=1999 etc. ;
proc format ;
value tf 1='1997' 2='1999' 3='2001' 4='2003' 5='2005' 6='2007' 7='2009'
8='2011' 9='2013' ;
value cf 0='No' 1='Yes' ;

```

```
run ;

title "Predicted Head's Wages by College Graduate Status and Time, Model
includes Random Statement and Results Combined by PROC MIANALYZE" ;
proc sgplot data=predicted_random ;
  series x=time y=predicted_hd wage_random / group=collegegrad ;
  xaxis type=discrete label='Time';
  yaxis label='Predicted Mean Head Wage/Salary 1997 to 2013 (in 2013
Dollars)' ;
  format collegegrad cf. time tf. ;
run ;
```

APPENDIX B – IVEWARE/SASMOD EXAMPLE

This example uses SAS-callable IVEware (v0.3) to repeat Example 2 using the SASMOD command along with built-in combining of multiply imputed data sets. The advantage of this tool lies in its ability to implement Jackknife Repeated Replication (Rust, 1985) for design-based variance estimation (in the absence of “PROC SURVEYMIXED” or equivalent in SAS) and to perform correct combining of multiply imputed data, all in one step. This makes the coding process a easier and also provides survey data analysts with needed variance estimation techniques.

In the following code, the OPTIONS statement sets the path and location of the IVEware software, the %SASMOD command invokes the SASMOD macro using PROC MIXED, the DATAIN statement reads each of 10 imputed data sets (LONG_IMPUTED1-LONG_IMPUTED10, saved previously by a DATA STEP process not shown here), the STRATUM, CLUSTER, and WEIGHT statements declare the complex sample variables and weight, and the PROC MIXED code is executed as previously presented in Example 2:

```
options set = SRCLIB 'E:\Srclib\sas' sasautos = (!SRCLIB' sasautos)
mautosource;
%sasmod (name=SASMOD_with_PROC_MIXED, dir=P:\SAS Global Forum 2018,
setup=new) ;
title "SASMOD with PROC MIXED, Random Intercept and Time" ;
datain w.long_imputed1 w.long_imputed2 w.long_imputed3 w.long_imputed4
w.long_imputed5 w.long_imputed6 w.long_imputed7 w.long_imputed8
w.long_imputed9 w.long_imputed10 ;
cluster cluster ;
stratum stratum ;
weight er34268 ;

/* SAS statements begin here */
proc mixed method=reml;
class collegegrad id ;
model headwage = time collegegrad time*collegegrad / solution ddfm=bw ;
random intercept time / type=un subject=id ;
run;
```

IVEware Output

IVEware Setup Checker, 05MAR18, 14:47:55

Setup listing:

```
title "SASMOD with PROC MIXED, Random Intercept and Time" ;
datain w.long_imputed1 w.long_imputed2 w.long_imputed3 w.long_imputed4
w.long_imputed5 w.long_imputed6 w.long_imputed7 w.long_imputed8 w.long_imputed9
w.long_imputed10 ;
cluster cluster ;
stratum stratum ;
weight er34268 ;
```

```
proc mixed method=reml;
class collegegrad id ;
model headwage = time collegegrad time*collegegrad / solution ddfm=bw ;
random intercept time / type=un subject=id ;
run;
```

IVEware Multiple Imputation Regression, Mon Mar 05 14:57:07 2018

1

"SASMOD with PROC MIXED, Random Intercept and Time"

All imputations

Valid cases 20403
Sum weights 885440.88

Degr freedom 783.3182835

-2 LogLike 517307.3438

Variable	Estimate	Std Error	Wald test	Prob > Chi
Intercept	71441.5999052	2059.0953257	1203.78642	0.00000
time _	-1678.0812797	400.4445230	17.56068	0.00003
collegegrad 0	-29797.7146714	2088.8503842	203.49383	0.00000
collegegrad 1	0.0000000	0.0000000	0.00000	0.00000
time*collegegrad 0	924.7157515	510.7795457	3.27755	0.07023
time*collegegrad 1	0.0000000	0.0000000	0.00000	0.00000

Variable	Estimate	95% Confidence Interval	
		Lower	Upper
Intercept	71441.5999052	67399.5948562	75483.6049542
time _	-1678.0812797	-2464.1540582	-892.0085011
collegegrad 0	-29797.7146714	-33898.1289135	-25697.3004293
collegegrad 1	0.0000000	0.0000000	0.0000000
time*collegegrad 0	924.7157515	-77.9447261	1927.3762292
time*collegegrad 1	0.0000000	0.0000000	0.0000000

Variable	Covariance Parameter
UN(1,1)	2341763994.35764
UN(2,1)	-127011229.71702
UN(2,2)	92999924.94601
Residual	132373713176.74280

The parameter estimates from both programs are identical but have slightly different standard errors, confidence intervals, Wald tests, and *p* values, as expected. The small differences in significance do not lead to different overall conclusions for this particular example. However, this will not always be the case when calculating correct design-based variances and thus, analyses that ignore complex sample design features can often lead to inaccurate conclusions.