Fractional Hot-Deck Imputation for Mixed Variables

Overview

Fractional hot-deck imputation (FHDI) (Kalton and Kish 1984; Fay 1996; Kim and Fuller 2004; Fuller and Kim 2005), also known as fractional imputation (FI), is a variation of hot-deck imputation in which one missing item for a recipient is imputed from multiple donors. Each donor donates a fraction of the original weight of the recipient such that the sum of the fractional weights from all the donors is equal to the original weight of the recipient. PROC SURVEYIMPUTE in SAS/STAT implements a FHDI method along with the fully efficient fractional imputation method, and some hot-deck imputation methods. For more information about fractional hot-deck imputation in SAS/STAT, see “The SURVEYIMPUTE Procedure”.

This example imputes missing values in both categorical and continuous variables by applying the FHDI method to a data set from the third National Health and Nutrition Examination Survey (NHANES III). The data set contains a set of BRR replicate weights. The REPWEIGHTS statement in PROC SURVEYIMPUTE accepts the BRR weights and creates imputation-adjusted replicate weights. The imputed data set and the imputation-adjusted replicate weights are then used in PROC SURVEYMEANS and PROC SURVEYREG to perform domain analysis and regression analysis, respectively.

The objective of NHANES is to study the health and nutritional status of the US population. NHANES uses a multistage stratified area sample with typically two PSUs per stratum. Strata are created on the basis of geographic location, metropolitan statistical area (MSA), and other demographic information. MSAs or a group of counties are selected as PSUs from each stratum. Sampling weights are unequal because of different selection probabilities among different subgroups and for reasons such as nonresponse and undercoverage. For more information about NHANES, see `http://www.cdc.gov/nchs/nhanes/about_nhanes.htm`.

NHANES III data contain missing values in many items. Multiple imputation was used to impute some of the missing items. Five multiply imputed data sets are available for public use. Because FHDI is used in this example to impute the missing values, you need the observed data, the missing (or imputation) flag for every item, and only one imputed data set. The data sets Core and IMP1 have been downloaded from the Centers for Disease Control and Prevention’s website (`https://www.cdc.gov/`). The Core data set contains the demographic variables, full sample weights, replicate weights, and imputation flags. The replicate weights are created by using Fay’s BRR method, with a Fay coefficient of 0.3. The IMP1 data set contains the first version of the five multiply imputed data sets.

The data set HealthMiss is obtained by merging Core and IMP1 data sets by the observation sequence number SEQN. The HealthMiss data set contains observation units that are between 17 and 60 years of age. Missing values are added according to the imputation flag in the Core data set. The following items are available in the HealthMiss data set for each observation unit:

- SEQN: observation sequence number
- WTPFQX6: observation weight, ranging from 220.18 to 140916.28

- WTPQRP1 to WTPQRP52: 52 replicate weights from the BRR method
- HSSEX: gender; 1 for male and 2 for female
- HFF1MI: anyone smokes cigarettes in the home; 1 for yes and 2 for no
- HAT28MI: activity level compared to others; –9 for not applicable, 1 for more active, 2 for less active, and 3 for about the same
- BMPHTMI: standing height (cm), ranging from 130.6 to 206.5
- BMPWTMI: body weight (kg), ranging from 26.75 to 241.80
- PEP6G3MI: K5 diastolic blood pressure (mmHg), ranging from 0 to 136
- HSAGEIR: age in years, ranging from 17 to 60
- HSHSIZER: household size; categories from 1 to 10
- Married: marital status; 1 for married and 0 for not married

There are no missing values in the variables SEQN, HSAGEIR, HSHSIZER, Married, and WTPFQX6 and in the replicate weight variables. The variables HFF1MI, HAT28MI, BMPHTMI, BMPWTMI, and PEP6G3MI contain missing values, which are imputed in this example. Although HFF1MI and HAT28MI have two and four observed levels, respectively, BMPHTMI, BMPWTMI, and PEP6G3MI have many observed levels. Because these last three variables have many observed levels, FHDI is applied to impute missing values jointly in all five variables.

Example: Imputation of Missing Values by Using FHDI

Before you apply the FHDI method to a data set, you should (1) incorporate auxiliary information in the imputation by creating imputation cells and (2) create bins for variables that have many levels.

Imputation cells divide the data into groups of similar units such that the recipient units have characteristics similar to those of the donor units in the same group. Characteristics of imputation cells might come from the same survey or from other sources, such as census data or previous surveys. The cell identification is known for every unit in the sample. For a helpful review, see Brick and Kalton (1996). For the purpose of this example, a cluster variable is created by using two demographic variables, HSAGEIR and HSHSIZER, and by using the FASTCLUS procedure in SAS/STAT. Both variables are available in the Core data set, and they do not contain missing values. These clusters are identified by the variable Cluster in the HealthMiss data set. Levels of the variables Cluster and Married are used to create imputation cells.

If you request FHDI, then the variables that have many levels (these are identified by the variables that you specify in the VAR statement but not in the CLASS statement) are first levelized to create bins. You can use the CLEVVAR= option to specify the variable that contains the bins for a numeric variable. Alternatively, you use the CLEVELS=*k* option to divide the observed range of the numeric variable into *k* equally spaced bins. First-stage FEFI is applied to the CLEVVAR= variables and to the variables that you specify in the CLASS statement. In this example, the bins are created for a numeric variable by using the observed 33rd and 66th percentiles for that variable. The CLEVVAR= variable should contain a missing value for every observation unit in which the corresponding numeric variable has a missing value. The following statements create the CLEVVAR= variables:

```
*---Create bins for continuous variables---;
data HealthMiss; set HealthMiss;
   if bmphtmi       = .     then bmphtlev=.;
   else if bmphtmi <= 162.6 then bmphtlev=1;
   else if bmphtmi <= 171.5 then bmphtlev=2;
   else                          bmphtlev=3;
   if bmpwtmi       = .    then bmpwtlev=.;
   else if bmpwtmi <= 65.7 then bmpwtlev=1;
   else if bmpwtmi <= 80.2 then bmpwtlev=2;
   else                         bmpwtlev=3;
   if pep6g3mi       = .    then pep6g3lev=.;
   else if pep6g3mi <= 65.3 then pep6g3lev=1;
   else if pep6g3mi <= 75.8 then pep6g3lev=2;
   else                          pep6g3lev=4;
   label bmphtlev  = "Bin values for BMPHTMI"
         bmpwtlev  = "Bin values for BMPWTMI"
         pep6g3lev = "Bin values for PEP6G3MI";
run;
```

The following statements use the FHDI method to impute the missing values:

```
proc surveyimpute data=HealthMiss method=fhdi varmethod=brr ndonors=5 seed=9388401;
   id seqn;
   class hff1mi hat28mi;
   weight wtpfqx6;
   repweights wtpqrp:;
   cells cluster married;
   var hff1mi hat28mi bmphtmi (clevvar=bmphtlev)
                      bmpwtmi (clevvar=bmpwtlev)
                      pep6g3mi(clevvar=pep6g3lev);
   output out=HealthFHDI;
run;
```

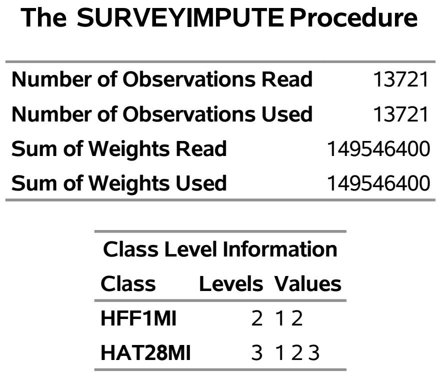
The PROC SURVEYIMPUTE statement invokes the procedure, the DATA= option specifies the input data set HealthMiss, the METHOD= option requests the FHDI method, the VARMETHOD= option requests the imputation-adjusted BRR replication weights, the NDONORS= options specifies the maximum number of second-stage donors, and the SEED= option specifies the random number generator seed. The variable SEQN in the ID statement identifies the observation units. The WEIGHT statement identifies the weight variable, and the REPWEIGHTS statement identifies the variables that contain the unadjusted BRR replicate weights. The CELLS statement identifies the imputation cell variables Cluster and Married, and the OUT= option in the OUTPUT statement names the output data set HealthFHDI.

The VAR statement specifies the variables in which the missing values are to be imputed. The CLASS statement identifies the categorical variables. Only first-stage FEFI will be applied to the CLASS variables. Second-stage FEFI and FHDI will be performed for the variables that are specified in the VAR statement but not in the CLASS statement. The CLEVVAR= option for these variables identifies the corresponding bin variables in which first-stage FEFI will be performed. In this example, you requested first-stage FEFI for the variables hff1mi, hat28mi, bmphtlev, bmpwtlev, and pep6g3lev; and second-stage FEFI and FHDI for the variables bmphtmi, bmpwtmi, and pep6g3mi.

You request that all five variables be imputed jointly and that the imputed data be saved in the HealthFHDI data set.

The number of observations and the CLASS level information are shown in Output 1. The Sum of Weights Read row shows that the 13,721 observation units in the sample represent over 149 million observation units in the population. The “Class Level Information” table displays the observed levels for the CLASS variables.

Figure 1 Imputation Information



The SURVEYIMPUTE Procedure

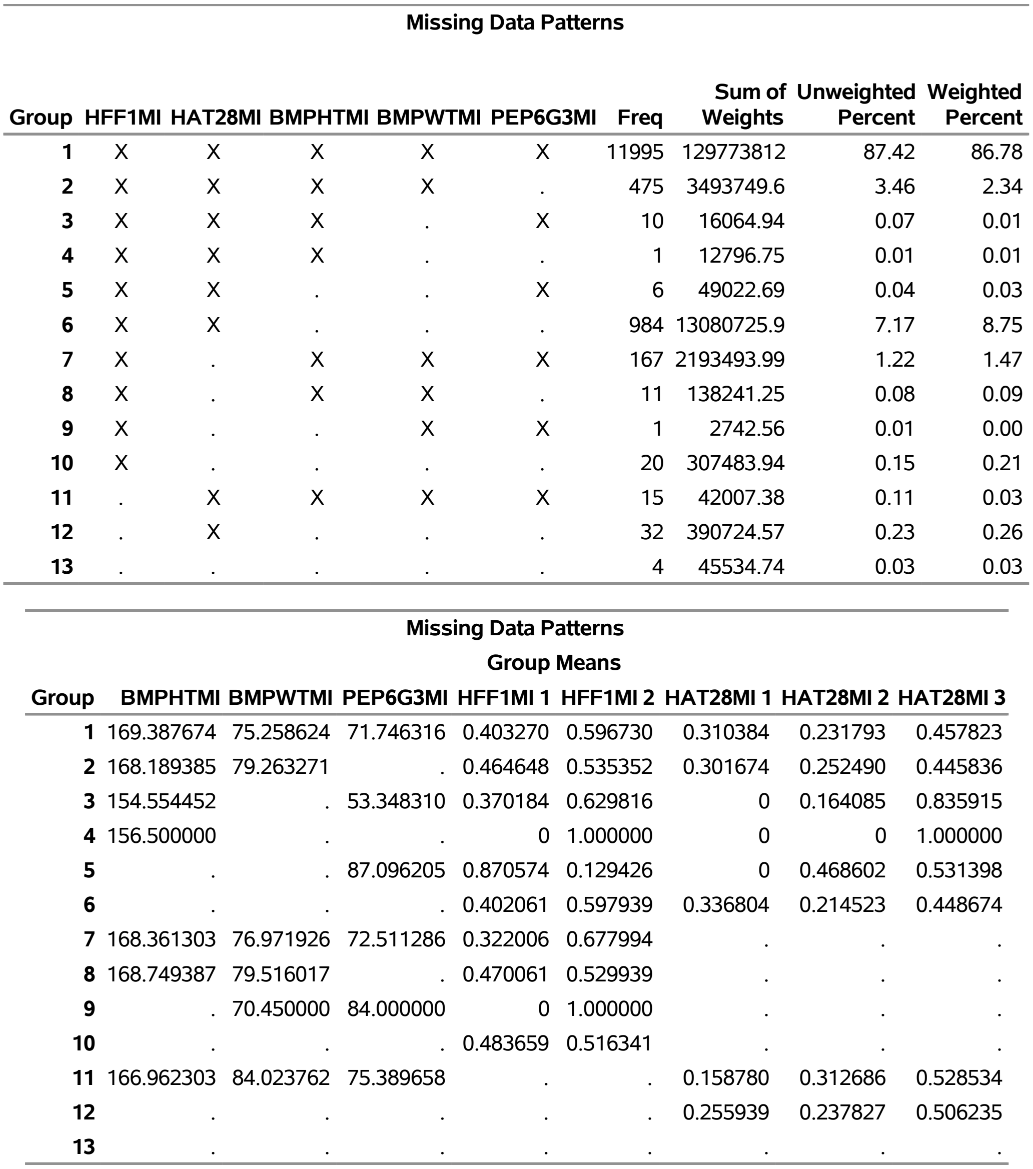
Number of Observations Read	13721
Number of Observations Used	13721
Sum of Weights Read	149546400
Sum of Weights Used	149546400

Class Level Information

Class	Levels	Values
HFF1MI	2	1 2
HAT28MI	3	1 2 3

The “Missing Data Patterns” table shows an arbitrary missing pattern. There are 13 different missing pattern groups. An “X” denotes that the variable is observed in that group, and a “.” denotes that the variable is missing. Almost 87.42% of the observation units have no missing values (Group 1), 7.17% of the observation units have missing values for the variables BMPHTMI, BMPWTMI, and PEP6G3MI (Group 6), and 0.03% of the observation units have missing values in all five variables (Group 13).

Figure 2 Missing Data Patterns



Missing Data Patterns

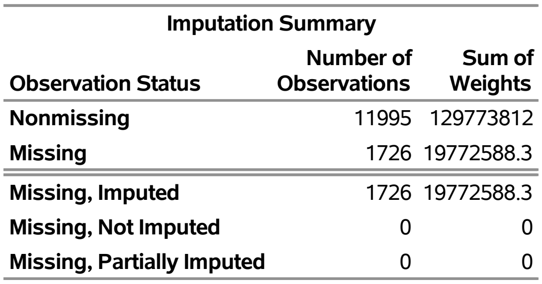
Group	HFF1MI	HAT28MI	BMPHTMI	BMPWTMI	PEP6G3MI	Freq	Sum of Weights	Unweighted Percent	Weighted Percent
1	X	X	X	X	X	11995	129773812	87.42	86.78
2	X	X	X	X	.	475	3493749.6	3.46	2.34
3	X	X	X	.	X	10	16064.94	0.07	0.01
4	X	X	X	.	.	1	12796.75	0.01	0.01
5	X	X	.	.	X	6	49022.69	0.04	0.03
6	X	X	.	.	.	984	13080725.9	7.17	8.75
7	X	.	X	X	X	167	2193493.99	1.22	1.47
8	X	.	X	X	.	11	138241.25	0.08	0.09
9	X	.	.	X	X	1	2742.56	0.01	0.00
10	X	.	.	.	.	20	307483.94	0.15	0.21
11	.	X	X	X	X	15	42007.38	0.11	0.03
12	.	X	.	.	.	32	390724.57	0.23	0.26
13	.	.	.	.	.	4	45534.74	0.03	0.03

Missing Data Patterns

	Group Means							
Group	BMPHTMI	BMPWTMI	PEP6G3MI	HFF1MI 1	HFF1MI 2	HAT28MI 1	HAT28MI 2	HAT28MI 3
1	169.387674	75.258624	71.746316	0.403270	0.596730	0.310384	0.231793	0.457823
2	168.189385	79.263271	.	0.464648	0.535352	0.301674	0.252490	0.445836
3	154.554452	.	53.348310	0.370184	0.629816	0	0.164085	0.835915
4	156.500000	.	.	0	1.000000	0	0	1.000000
5	.	.	87.096205	0.870574	0.129426	0	0.468602	0.531398
6	.	.	.	0.402061	0.597939	0.336804	0.214523	0.448674
7	168.361303	76.971926	72.511286	0.322006	0.677994	.	.	.
8	168.749387	79.516017	.	0.470061	0.529939	.	.	.
9	.	70.450000	84.000000	0	1.000000	.	.	.
10	.	.	.	0.483659	0.516341	.	.	.
11	166.962303	84.023762	75.389658	.	.	0.158780	0.312686	0.528534
12	.	.	.	.	.	0.255939	0.237827	0.506235
13	.	.	.	.	.	.	.	.

The “Imputation Summary” table in Output 3 displays the number of observation units without any missing items (11,995), the number of observation units that contain at least one missing item (1,726), and the number of units in which the missing values are imputed. Missing values in all 1,726 units are imputed.

Figure 3 Imputation Summary



Imputation Summary

Observation Status	Number of Observations	Sum of Weights
Nonmissing	11995	129773812
Missing	1726	19772588.3
Missing, Imputed	1726	19772588.3
Missing, Not Imputed	0	0
Missing, Partially Imputed	0	0

Because fractional imputation replaces one observed unit with several observation rows that contain imputed values, the 13,721 observed units in the input data set HealthMiss generates 127,002 observation rows in the imputed data set HealthFHDI. The following note displays the number of observation rows (127,007) in the imputed data set HealthFHDI:

```
NOTE: The data set WORK.HEALTHFHDI has 127007 observations and 134 variables.
```

Example: Analysis for Fractionally Imputed Data

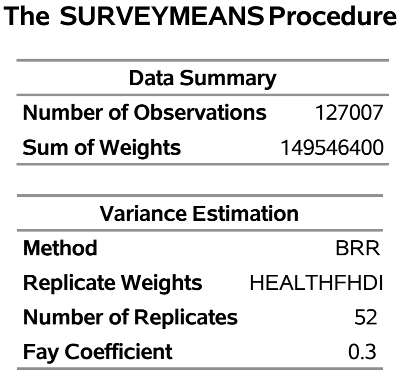
You can use the imputed data set, the imputation-adjusted replicate weights, and the appropriate Fay coefficient to compute any estimators from your imputed data. However, you must use the REPWEIGHTS statement in SAS/STAT survey analysis procedures to specify the imputation-adjusted replicate weights. The following two examples describe a domain analysis and a regression analysis that use the imputed data.

The following PROC SURVEYMEANS statements estimate the mean diastolic blood pressure in the year 2000 population, and in the subpopulation of smokers and nonsmokers.

```
ods graphics on;
proc surveymeans data=HealthFHDI varmethod=brr(Fay=0.3) plots=domain;
   weight ImpWt;
   repweights ImpRepWt_:;
   var pep6g3mi;
   domain hff1mi;
run;
```

The “Data Summary” table in Output 4 displays the number of observation rows (127,007) and the sum of weights (149,546,400). Because fractional imputation is used, the number of observation rows is not equal to the number of observation units (13,721). However, the sum of weights from the observation rows, which is an estimate of the population size, is the same as the sum of weights from the observation units. The “Variance Estimation” table in Output 4 shows that Fay’s BRR with 52 replicate weights and a Fay coefficient of 0.3 are used for variance estimation.

Figure 4 Data Summary



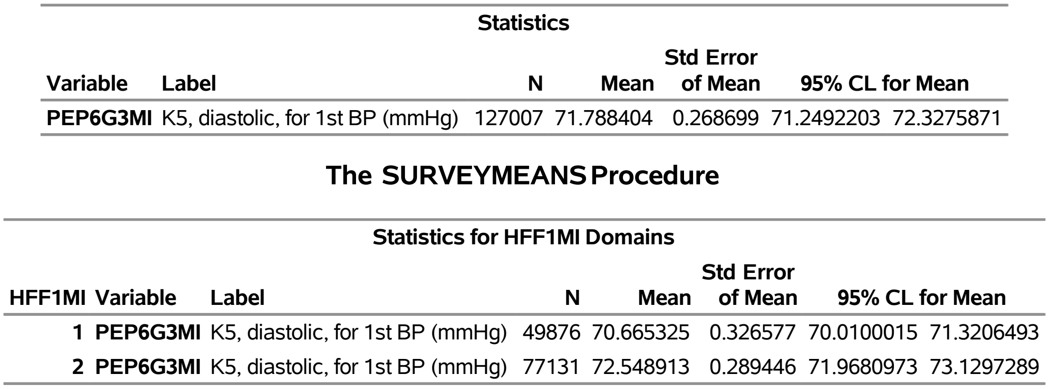
The SURVEYMEANS Procedure

Data Summary	
Number of Observations	127007
Sum of Weights	149546400

Variance Estimation	
Method	BRR
Replicate Weights	HEALTHFHDI
Number of Replicates	52
Fay Coefficient	0.3

The “Mean Diastolic Blood Pressure” table in Output 5 displays the mean diastolic blood pressure for the overall population as 71.78 with a standard error of 1.34. For smokers the mean diastolic blood pressure is 70.67 with a standard error of 1.63, and for nonsmokers it is 72.55 with a standard error of 1.45. The “N” column displays the number of observation rows, not the number of observation units.

Figure 5 Mean Diastolic Blood Pressure



Statistics

Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
PEP6G3MI	K5, diastolic, for 1st BP (mmHg)	127007	71.788404	0.268699	71.2492203	72.3275871

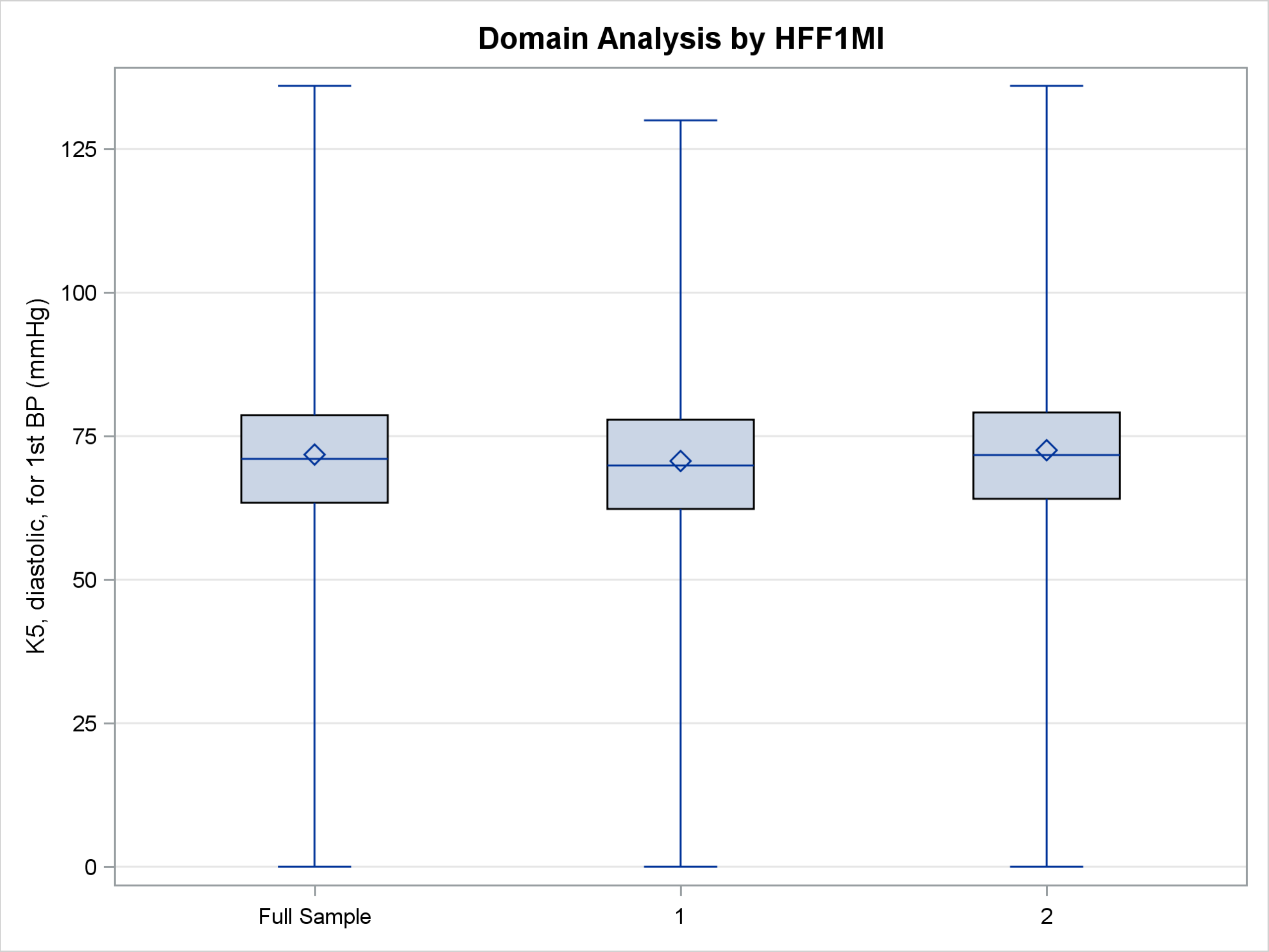
The SURVEYMEANS Procedure

Statistics for HFF1MI Domains

HFF1MI	Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
1	**PEP6G3MI**	K5, diastolic, for 1st BP (mmHg)	49876	70.665325	0.326577	70.0100015	71.3206493
2	**PEP6G3MI**	K5, diastolic, for 1st BP (mmHg)	77131	72.548913	0.289446	71.9680973	73.1297289

A box plot of the weighted distribution of diastolic blood pressure is displayed in Output 6. The first box is for the overall population, and the other two boxes are for the two domains defined by smoking habits.

Figure 6 Diastolic Blood Pressure

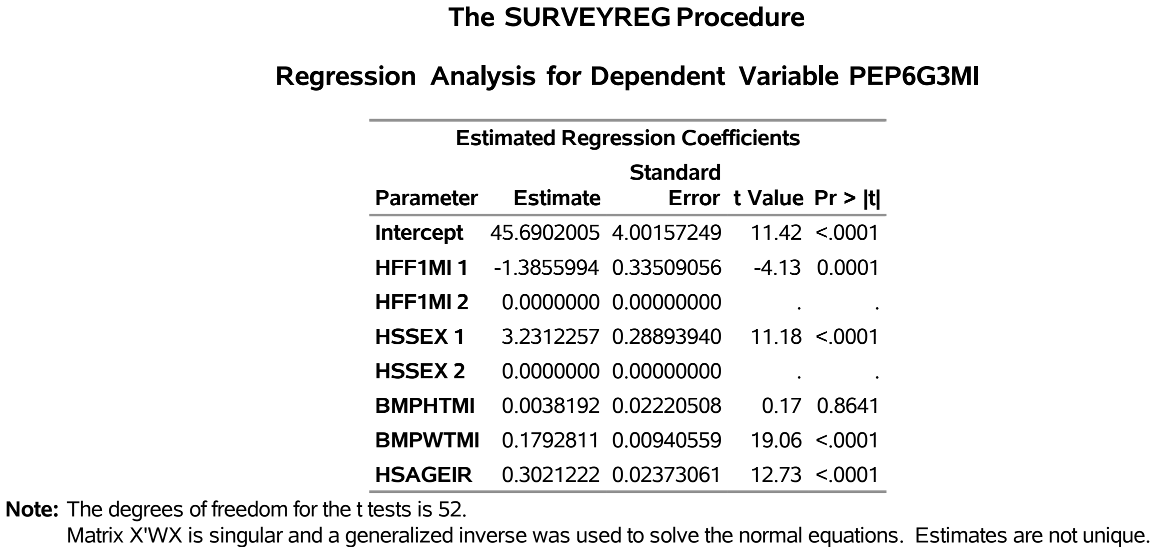


The following PROC SURVEYREG statements estimate the regression coefficients for regressing diastolic blood pressure on smoking status, gender, height, weight, and age. Imputation-adjusted weights and imputation-adjusted replicate weights are used for point estimation and variance estimation, respectively. The SOLUTION option in the MODEL statement displays the parameter estimates. Estimated values for the regression parameters along with their standard errors are shown in Output 7. The OUT= option in the OUTPUT statement saves the residuals and the fitted values in the SAS data set Resid.

```
proc surveyreg data=HealthFHDI varmethod=brr(Fay=0.3);
   weight ImpWt;
   repweights ImpRepWt_:;
   class hff1mi hssex;
   model pep6g3mi = hff1mi hssex bmphtmi bmpwtmi hsageir / solution;
   output out=Resid residual=Residuals predicted=Fitted;
run;
```

Estimated regression parameters and their standard errors are displayed in Output 7. All covariates except height (BMPHTMI) have small standard errors compared to their estimated values. Thus, they are all important in describing the regression relationship in the NHANES III population between 17 and 60 years of age. The degrees of freedom for the *t* tests is 52, which is equal to the number of BRR replicates.

Figure 7 Parameter Estimates



The SURVEYREG Procedure

Regression Analysis for Dependent Variable PEP6G3MI

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	45.6902005	4.00157249	11.42	<.0001
HFF1MI 1	-1.3855994	0.33509056	-4.13	0.0001
HFF1MI 2	0.0000000	0.00000000	.	.
HSSEX 1	3.2312257	0.28893940	11.18	<.0001
HSSEX 2	0.0000000	0.00000000	.	.
BMPHTMI	0.0038192	0.02220508	0.17	0.8641
BMPWTMI	0.1792811	0.00940559	19.06	<.0001
HSAGEIR	0.3021222	0.02373061	12.73	<.0001

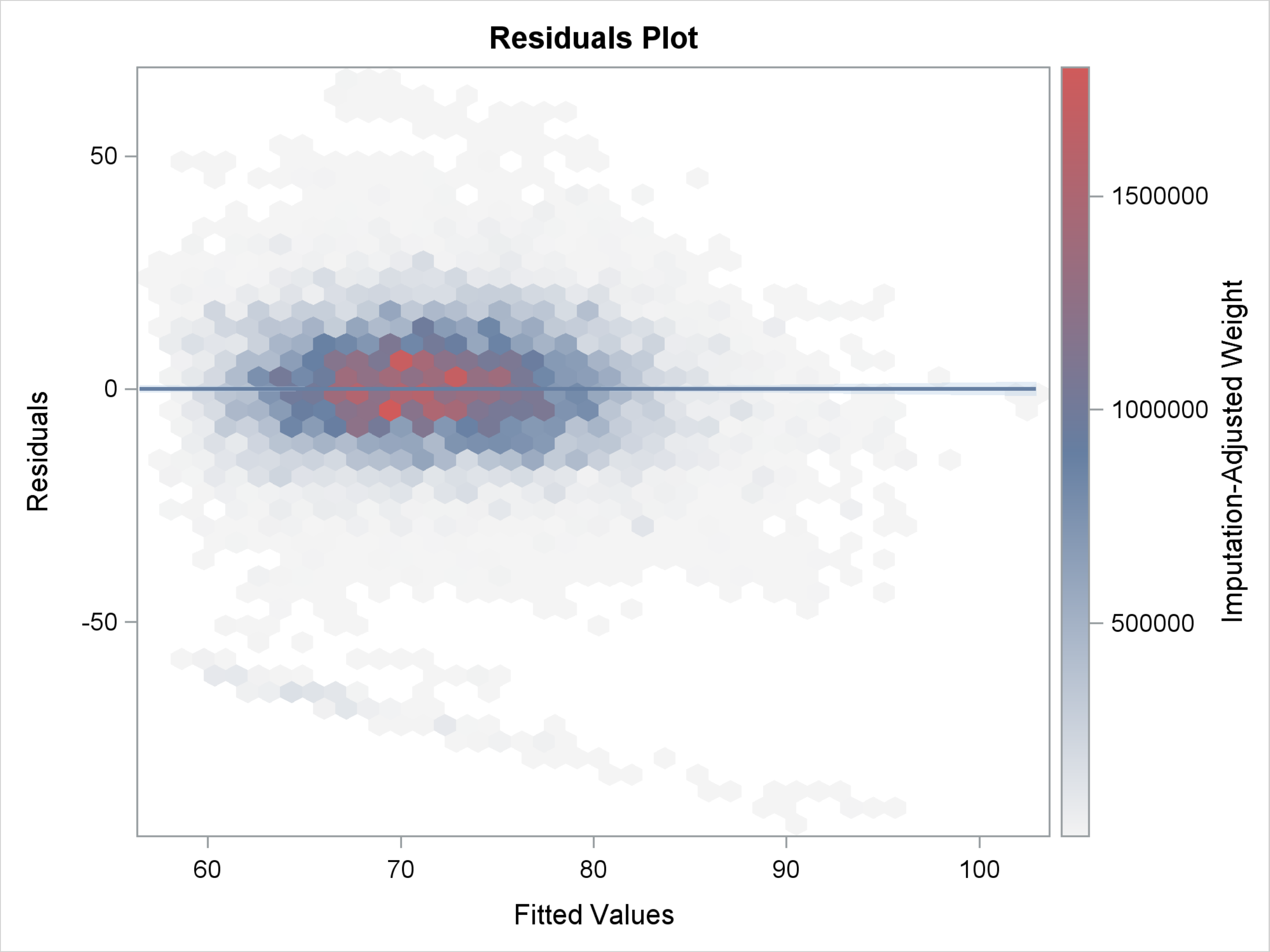
Note: The degrees of freedom for the t tests is 52.
Matrix X'WX is singular and a generalized inverse was used to solve the normal equations. Estimates are not unique.

The following PROC SURVEYREG statements request a residual plot by using the residuals and fitted values from the Resid data set. Imputation-adjusted weights are displayed by using a heat map, as shown in Output 8. The SHAPE=HEXAGONAL option requests hexagonal bins, and the NBINS=60 option specifies that 60 bins be used. For more information about how to create customized graphs by using ODS graphs, see Kuhfeld (2016). The weighted residual plot does not reveal any major violations from the model assumptions—namely, linearity and equal variance.

Because fractional imputation increases the number of observation rows, you must use imputation-adjusted weights in all analyses that involve fractionally imputed data sets.

```
ods graphics on;
proc surveyreg data=resid plots(nbins=60)=fit(shape=hexagonal);
   model Residuals=Fitted;
   weight impwt;
run;
```

Figure 8 Plot of Residuals versus Fitted Values



References

Brick, J. M., and Kalton, G. (1996). "Handling Missing Data in Survey Research." *Statistical Methods in Medical Research* 5:215–238.

Fay, R. E. (1996). "Alternative Paradigms for the Analysis of Imputed Survey Data." *Journal of the American Statistical Association* 91:490–498.

Fuller, W. A., and Kim, J. K. (2005). "Hot Deck Imputation for the Response Model." *Survey Methodology* 31:139–149.

Kalton, G., and Kish, L. (1984). "Some Efficient Random Imputation Methods." *Communications in Statistics—Theory and Methods* 13:1919–1939.

Kim, J. K., and Fuller, W. A. (2004). "Fractional Hot Deck Imputation." *Biometrika* 91:559–578.

Kuhfeld, W. F. (2016). "Highly Customized Graphs Using ODS Graphics." In *Proceedings of the SAS Global Forum 2016 Conference*. Cary, NC: SAS Institute Inc. `http://support.sas.com/resources/papers/proceedings16/SAS1800-2016.pdf`.