Imputation Methods for Surveys: A Demonstration of the IMPUTE procedure in SUDAAN

Kimberly Ault, RTI International, RTP, NC

Abstract

Survey researchers commonly encounter missing data during analysis. Ad-hoc missing data methods, such as complete-case analysis, are easy to implement but they have well-known disadvantages of potentially yielding biased results and having reduced power due to deleting observations with missing values. The impact of missing data on survey estimates depends on the pattern of missing data, percent of missing data, and parameters to be estimated. Since most surveys experience some missing data, survey data analysis should account for missing data. Weighting adjustments may compensate for non-coverage and unit nonresponse, but imputation methods that assign values for missing responses are more commonly used to compensate for item nonresponse. The IMPUTE Procedure in SUDAAN v11 performs the following imputation methods: weighted sequential hot deck imputation, cell mean imputation, linear regression for continuous variables, and logistic regression for binary variables. Data from public use files for the 1997-2004 National Health Interview Survey linked to the National Death Index are used to illustrate each imputation method. Advantages and disadvantages are discussed in the summary section.

Analysis Goals

- Assess the association serious psychological distress (SPD) at the time of interview and mortality.
- Determine if SPD is a significant risk factor for mortality even after controlling for sociodemographic characteristics and other behavioral risk factors.

Data Source

- National Health Interview Survey (NHIS) public use file from 1997 to 2004
- National Death Index (NDI) public use files from 1997 to 2006
- From 1997 to 2004, there were 258,279 adult respondents in the NHIS
- 15,882 records (approximately 6 percent of the total records) ineligible for linkage to NDI
- Eligible sample of 242,397 NHIS respondents aged 18 or older

Ten variables from the 2004 NHIS data file were considered in the analysis. The variable type and amount of missing data for these variables are as shown in Table 1.

Table 1. Distribution of Variables Requiring Imputation

Variable	Variable Type	Variable Values	Unweighted Number of Non-Missing	Unweighted Number of Missing	Unweighted Percent Missing	Weighted Percent Missing
Age at Interview	Continuous	18-99	0	0	NA	NA
Gender	Nominal	• Male • Female	0	0	NA	NA
Race/Ethnicity	Nominal	• Hispanic • Non-Hispanic White • Non-Hispanic Black • Non-Hispanic Other	0	0	NA	NA
Marital Status	Nominal	• Married • Separated or Divorced • Widowed • Never married	241,795	602	0.25	0.18
Height	Continuous	59 to 76 inches	225,924	16,473	6.80	6.52
Weight	Continuous	99 to 285 pounds	221,345	21,052	8.68	8.49
Education	Ordinal	• Less than High School • High School Graduate/General Equivalency Diploma/Some College • College Graduate	240,899	1,498	0.62	0.62
Smoking Status	Nominal	• Daily smoker • Occasional smoker • Former smoker • Never Smoked	241,016	1,381	0.57	0.54
Number of Chronic Conditions	Ordinal	• None • One • Two or More	240,293	2,104	0.87	0.76
SPD	Dichotomous	• Yes • No	238,541	3,856	1.48	1.47

NA=Not Applicable. Source: 1997 to 2004 CDC/NCHS National Health Interview Survey.

Imputation Methods

Four imputation methods were used and were dependent upon the variable type:

- Weighted Sequential Hot Deck for Categorical Variables (Nominal and Ordinal)
- Linear Regression for Continuous Variables
- Cell Mean for Continuous Variables
- Logistic Regression for Binary Variables

Example 1 – Weighted Sequential Hot Deck Imputation for Categorical Variables

Weighted Sequential Hot Deck Imputation: Sequential hot-deck imputation is a common method used for item nonresponse in survey research. This method uses the respondent survey data (donors) to provide imputed values for records with missing values by defining imputation classes, which generally consist of a cross-classification of covariates, and then replacing the missing values with the randomly selected donor values within the imputation classes. When sequential hot-deck imputation is performed using the sampling weights associated with the survey, the method is called *weighted* sequential hot-deck imputation (WSHD).

Smoking status, chronic indicator, education, BMI, and marital status were imputed using the **WSHD option (method=WSHD)** in the IMPUTE procedure.

SUDAAN CODE:

```
proc impute data=in method=wshd;
  weight wt8;
  class smoke chronic marital educ_cat;
  impby agegrp sex racehisp;
  impvar smoke chronic marital educ_cat;
  impid numpublicid;
  print;
```

Imputation Classes: Age, Gender, Race/Ethnicity

Table 2. Results from Example 1- Before and After Imputation Percentages

Variable	Before Imputation	After Imputation	Absolute Difference	Relative Difference (%)
Education	Weighted Percent (SE)			
Less than High School	17.79 (0.19)	17.82 (0.19)	0.03	0.17
High School Graduate/General Equivalency Diploma/Some College	58.83 (0.20)	58.82 (0.20)	0.01	-0.02
College Graduate	23.39 (0.24)	23.36 (0.24)	0.03	-0.13
Marital Status	Weighted Percent (SE)			
Married	63.95 (0.21)	63.96 (0.21)	0.01	0.02
Separated or Divorced	10.47 (0.08)	10.47 (0.08)	0.00	0.00
Widowed	6.62 (0.08)	6.62 (0.08)	0.00	0.00
Never married	18.95 (0.20)	18.95(0.20)	0.00	0.00
Smoking Status	Weighted Percent (SE)			
Daily smoker	18.97 (0.15)	18.96 (0.15)	0.01	-0.05
Occasional smoker	4.20 (0.05)	4.20 (0.05)	0.00	0.00
Former smoker	22.57 (0.13)	22.57 (0.13)	0.00	0.00
Never Smoked	54.26 (0.18)	54.26 (0.18)	0.00	0.00
Chronic Conditions	Weighted Percent (SE)			
None	60.58 (0.16)	60.46 (0.16)	0.12	-0.20
One	24.77 (0.11)	24.80 (0.11)	0.03	0.12
Two or More	14.66 (0.12)	14.74 (0.12)	0.08	0.55

SE=Standard Error

Note: The absolute difference is the absolute difference of the post-imputation and the pre-imputation percentages. The relative percent difference is defined as 100 * [(post-imputation percentage – pre-imputation percentage) / pre-imputation percentage].

Source: 1997 to 2004 CDC/NCHS National Health Interview Survey.

Example 2 - Linear Regression for Continuous Variables

Linear Regression Imputation: Linear regression imputation is the process of replacing missing values with a predicted or expected value computed from a fitted regression model. The IMPUTE procedure obtains the predicted values from the fitted linear regression models and then replaces the missing values for the variables.

The height and weight variables were imputed using the **linear regression option (method=linear)** in the IMPUTE procedure and then the imputed values were used to compute a body mass index value that was used in the Cox proportional hazard models.

SUDAAN CODE:

```
proc impute data=in method=linear;
   weight wt8;
   class agegrp sex racehisp;
   impmodel agegrp sex racehisp;
   impvar wt ht;
   impid numpublicid;
  print;
```

Models: Weight = Age, Gender, Race/Ethnicity, Height = Age, Gender, Race/Ethnicity

Table 3. Results from Example 2 – Before and After Imputation Means

Variable	Before Imputation		After Imputation		Absolute Difference	Relative Difference (%)
	Mean	SE	Mean	SE		
Weight	169.62	0.1118	169.28	0.1039	0.33	-0.20
Height	66.99	0.0115	66.98	0.0110	0.01	-0.02

SE=Standard Error

Note: The absolute difference is the absolute difference of the post-imputation and the pre-imputation means. The relative percent difference is defined as 100 * [(post-imputation mean – pre-imputation mean) / pre-imputation mean].

Source: 1997 to 2004 CDC/NCHS National Health Interview Survey.

Table 4. Results from Example 2 – Before and After Imputation Percentages

BMI	Computed based on imputed Height and Weight			
	Percentage (SE)		Absolute Difference	Relative Difference (%)
	Before Imputation	After Imputation		
BMI < 20 = Underweight	5.78 (0.06)	5.30 (0.06)	0.48	-8.30
20 ≤ BMI < 25 = Normal Weight	38.85 (0.15)	35.68 (0.14)	3.17	-8.16
25 ≤ BMI < 30 = Overweight	36.59 (0.13)	39.89 (0.12)	3.30	9.02
BMI ≥ 30 = Obese	20.78 (0.12)	19.13 (0.11)	1.65	-7.94

SE=Standard Error

Note: The absolute difference is the absolute difference of the post-imputation and the pre-imputation percentages. The relative percent difference is defined as 100 * [(post-imputation percentage – pre-imputation percentage) / pre-imputation percentage].

Source: 1997 to 2004 CDC/NCHS National Health Interview Survey.

Example 3 - Cell Mean Imputation for Continuous Variables

Cell Mean Imputation- Cell mean imputation is the process of replacing missing values with the mean value computed within a group of respondents (or imputation cell).

The height and weight variables were imputed using the **cell mean option (method=cellmn)** in the IMPUTE procedure and then the imputed values were used to compute a body mass index value that was used in the Cox proportional hazard models.

SUDAAN CODE:

```
proc impute data=in method=cellmn;
  weight wt8;
  impby agegrp sex racehisp;
  impvar wt ht;
  impid numpublicid;
  print;
```

Imputation Classes: Age, Gender, Race/Ethnicity

Table 5. Results from Example 3 – Before and After Imputation Means

	Before Imputation		After Imputation			
Variable	**Mean**	**SE**	**Mean**	**SE**	**Absolute Difference**	**Relative Difference (%)**
Weight	169.62	0.1118	169.28	0.1040	0.34	-0.20
Height	66.99	0.0115	66.98	0.0110	0.01	-0.02

SE=Standard Error

Note: The absolute difference is the absolute difference of the post-imputation and the pre-imputation means. The relative percent difference is defined as 100 * [(post-imputation mean – pre-imputation mean) / pre-imputation mean].

Source: 1997 to 2004 CDC/NCHS National Health Interview Survey.

Table 6. Results from Example 3 – Before and After Imputation Percentages

	Computed based on imputed Height and Weight			
	Percentage (SE)			
BMI	**Before Imputation**	**After Imputation**	**Absolute Difference**	**Relative Difference (%)**
BMI < 20 = Underweight	5.78 (0.06)	5.30 (0.06)	0.48	-8.30
20 ≤ BMI < 25 = Normal Weight	38.85 (0.15)	35.21 (0.14)	3.64	-9.37
25 ≤ BMI < 30 = Overweight	36.59 (0.13)	40.36 (0.13)	3.77	10.30
BMI ≥ 30 = Obese	20.78 (0.12)	19.14 (0.11)	1.64	-7.89

SE=Standard Error

Note: The absolute difference is the absolute difference of the post-imputation and the pre-imputation percentages. The relative percent difference is defined as 100 * [(post-imputation percentage – pre-imputation percentage) / pre-imputation percentage].

Source: 1997 to 2004 CDC/NCHS National Health Interview Survey.

Example 4 – Logistic Regression for Binary Variables

Logistic Regression Imputation: Logistic regression imputation is the process of replacing missing values with a predicted or expected value computed from a regression model. The IMPUTE procedure obtains the predicted values from the fitted logistic regression model. In addition, each item nonrespondent record has a random number assigned to it from a uniform distribution and the predicted values are then compared to the random number to determine the final imputed value.

The dichotomous serious psychological distress (SPD) variable was imputed using the **logistic regression option (method=logistic)** in the IMPUTE procedure.

SUDAAN CODE:

```
proc impute data=in method=logistic;
  weight wt8;
  class agegrp sex racehisp;
  impmodel agegrp sex racehisp;
  impvar spd2;
  impid numpublicid;
  print;
```

Model SPD= Age, Gender, Race/Ethnicity

Table 7. Results from Example 4 – Before and After Imputation Percentages

	Before Imputation		After Imputation		Absolute	Relative
SPD2	**Percentage**	**SE**	**Percentage**	**SE**	**Difference**	**Difference**
No	96.93	0.0532	96.94	0.0524	0.01	0.01
Yes	3.07	0.0532	3.06	0.0524	0.01	-0.33

SE=Standard Error

Note: The absolute difference is the absolute difference of the post-imputation and the pre-imputation percentages. The relative percent difference is defined as 100 * [(post-imputation percentage – pre-imputation percentage) / pre-imputation percentage].

Source: 1997 to 2004 CDC/NCHS National Health Interview Survey.

Comparison of Imputation Methods

The variables requiring imputation were imputed by modeling the relationships between the variables without missing data and each of the variables with missing values. The Cox proportional hazard regression was used to model the survival time and the same set of covariates with and without imputation. Results from the using different imputation methods are displayed in Table 8.

Method 1: No Imputation

Method2: Impute Marital Status, Education Level, Smoking Status, Number of Chronic Conditions (WSHD), Impute Height and Weight (Linear Regression), Compute BMI, Impute SPD2 (Logistic Regression)

Method 3: Impute Marital Status, Education Level, Smoking Status, Number of Chronic Conditions, Impute Height and Weight (Cell Mean), Compute BMI, Impute SPD2 (Logistic Regression)

Method4: Impute Marital Status, Education Level, Smoking Status, Number of Chronic Conditions (WSHD), Compute BMI, Impute BMI (WSHD), Impute SPD2 (Logistic Regression)

Table 8: Proportional Hazards Regression
Relationship between SPD and Time to Death Adjusted for Age, Gender, Race/Hispanicity, Marital Status, BMI, Education, Smoking Status, and Number of Chronic Conditions for Persons 35 to 44 years, NHIS 1997 - 2004

	Method 1	**Method 2**	**Method 3**	**Method 4**
Variable	**Hazard Ratio (CI)**	**Hazard Ratio (CI)**	**Hazard Ratio (CI)**	**Hazard Ratio (CI)**
SPD2				
No	1.00	1.00	1.00	1.00
Yes	1.25 (0.90,1.73)	1.37 (1.00,1.87)	1.37 (1.00,1.87)	1.37 (1.00,1.87)
Wald P	*0.2042*	*0.0484*	*0.0484*	*0.0543*
BMI				
Underweight	0.95 (0.85,1.06)	1.67 (1.13,2.48)	1.68 (1.13,2.49)	1.47 (1.00,2.17)
Normal Range	1.00	1.00	1.00	1.00
Overweight	1.71 (1.40,2.10)	1.17 (0.97,1.42)	1.18 (0.97,1.43)	1.05 (0.85,1.29)
Obese	1.74 (0.91, 3.32)	1.07 (0.83,1.39)	1.07 (0.83,1.39)	1.10 (0.86,1.40)
Wald P	*0.0004*	*0.0513*	*0.0474*	*0.0513*

Source: 1997 to 2004 CDC/NCHS National Health Interview Survey.

For ***Method 1 (No Imputation)*** the hazard ratio for SPD is 1.25 (implying an 25% increase in hazard) and indicates that although death is observed to occur sooner for adults aged 35 to 44 with SPD there is no statistically significant association, since the 95% confidence interval contains the null value of 1.0. Additionally for ***Method 1 (No Imputation)***, the Wald p-value (0.2042) for testing main effects model shows that SPD is not significantly associated with follow-up time to death.

For ***Method 2 (Imputing height and weight with linear regression)*** *and* ***Method 3 (Imputing height and weight with cell mean),*** the hazard ratio increases to 1.37 and that SPD is statistically significant, since the 95% confidence interval contains the null value of 1.0. The Wald p-values support this conclusion with values less than 0.05.

For ***Method 4 (Computing BMI from height and weight and then imputing BMI with WSHD),*** the Wald p-value is slightly above 0.05 (0.0543).

For ***Methods 1 and 3***, BMI is significantly associated with follow-up time to death. However, for ***Methods 2 and 4*** the p-values increase to slightly larger than 0.05.

Advantages

Weighted Sequential Hot Deck

- Uses actual values from the data.
- Uses sample weights in the imputation process.
- Preserves the weighted distribution of post-imputation variables across imputation classes when compared to the weighted distribution of pre-imputation variables.

Cell Mean

- Provides an unbiased estimate of the overall variable mean if the probability of nonresponse is the same for every respondent in a class or if values within a class are not related to the probability of nonresponse.

Linear and Logistic Regression Imputation

- Models can include a large number of variables to capture data relationships.

Disadvantages

Weighed Sequential Hot Deck

- Fails to capture multivariate relationships.

Cell Mean

- Weakens covariance and correlation between variables since relationships between variables are ignored.

Linear and Logistic Regression Imputation

- Sensitive to model misspecification.
- Assumes that the same model explains the data for the non-missing cases as for the missing cases, which of course in not necessarily true

Summary

- Case deletion strategies assume that the deleted cases are a relatively small proportion of the entire dataset and that the complete cases are a representative sample.
- Loss in sample size can appreciably diminish the statistical power of the analysis.
- As a rule of thumb, if a variable has more than 5% missing values, cases are not deleted, and many researchers are much more stringent than this.
- Many different approaches to imputation and imputation method should be based on the goal of the analysis.
- Other factors that are important in determining the type of imputation method include: 1) size of data file, 2) level of missingness of data, and 3) patterns of missing data, or 4) structure of the data such as cross-sectional or longitudinal data.

References

National Center for Health Statistics, Office of Analysis and Epidemiology. (2009, December 2). *The National Health Interview Survey (1986-2004) linked mortality files, mortality follow-up through 2006: Matching methodology*. Retrieved from http://www.cdc.gov/nchs/data/datalinkage/matching_methodology_nhis_final.pdf

National Center for Health Statistics (1994 -2004). *Data File Documentation, National Health Interview Survey*. National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland. http://www.cdc.gov/nchs/nhis/quest_data_related_1997_forward.htm

RTI International. (2012). *SUDAAN® Language Manual, Release 11.0*. Research Triangle Park, NC: RTI International.

Kimberly Ault, Ph.D.
RTI International
Research Statistician
3040 Cornwallis Rd
Cox Building Room 230
Research Triangle Park, NC 27709
919-541-7455
ault@rti.org