Journal of Modern Applied Statistical Methods

Volume 16 | Issue 1 Article 39

5-1-2017

Using Multiple Imputation to Address Missing Values of Hierarchical Data

Yujia Zhang
Centers for Disease Control and Prevention, Atlanta, coi8@cdc.gov

Sara Crawford
Centers for Disease Control and Prevention, Atlanta, sgv0@cdc.gov

Sheree Boulet
Centers for Disease Control and Prevention, Atlanta, sbu1@cdc.gov

Michael Monsour
Centers for Disease Control and Prevention, Atlanta, mhm2@cdc.gov

Bruce Cohen
Massachusetts Department of Public Health, Boston, bruce.cohen@state.ma.us

See next page for additional authors

Recommended Citation
Zhang, Y., Crawford, S., Boulet, S., Monsour, M., Cohen, B., McKane, P., & Freeman, K. (2017). Using multiple imputation to address missing values of hierarchical data. Journal of Modern Applied Statistical Methods, 16(1), 744-752. doi: 10.22237/jmasm/1493599140

Using Multiple Imputation to Address Missing Values of Hierarchical Data

Cover Page Footnote

Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Authors

Yujia Zhang, Sara Crawford, Sheree Boulet, Michael Monsour, Bruce Cohen, Patricia McKane, and Karen Freeman

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 744-752.
doi: 10.22237/jmasm/1493599140

ISSN 1538 − 9472

Using Multiple Imputation to Address Missing Values of Hierarchical Data

Yujia Zhang
Centers for Disease Control and Prevention, Atlanta, GA

Sara Crawford
Centers for Disease Control and Prevention, Atlanta, GA

Sheree Boulet
Centers for Disease Control and Prevention, Atlanta, GA

Michael Monsour
Centers for Disease Control and Prevention, Atlanta, GA

Bruce Cohen
MA Department of Health
Boston, MA

Patricia McKane
MI Dept. of Comm. Health
Lansing, MI

Karen Freeman
FL Department of Health
Tallahassee, FL

Missing data may be a concern for data analysis. If it has a hierarchical or nested structure, the SUDAAN package can be used for multiple imputation. This is illustrated with birth certificate data that was linked to the Centers for Disease Control and Prevention's National Assisted Reproductive Technology Surveillance System database. The Cox-Iannacchione weighted sequential hot deck method was used to conduct multiple imputation for missing/unknown values of covariates in a logistic model.

Keywords: Hierarchical or nesting structure, multiple imputation, weighted sequential hot deck

Introduction

Population-based hierarchical or nested data and multiple covariates are often used in maternal and child health research. The covariates may contain unknown/missing values, which are excluded in traditional model fitting such that only complete cases are used. Although the percent of unknown/missing values for one variable is usually small, the percent of unknown/missing values across all covariates may be larger. Using only complete cases in analysis reduces the effective sample size and testing power, which is especially concerning when the

Yujia Zhang is a Mathematic Statistician in the Division of Reproductive Health, National Center for Chronic Disease Prevention and Health Promotion, CDC. Email at coi8@cdc.gov.

outcome is infrequent since it likely introduces small-sample bias in logistic model fitting (King & Zeng, 2001; Rotnitzky & Wypij, 1994).

One strategy to address the impact of missing values on parameter estimates is to use imputed data in analysis. A single imputation method fills each missing entry with an imputed value, such that standard complete-data methods can be used for analysis. This method ignores the variability contributed by the lack of information on the missing values, leading to variance underestimation. Another method, multiple imputation replaces each missing entry with two or more values and draws inferences by combining the results of several complete-data analyses to address within and between-imputation variability in variance estimation (Rubin, 1986, 1997; Schafer, 1999).

The traditional multiple imputation method used by most commercial statistical software packages such as SAS, IVEware, etc., adopts a parametric approach such as regression imputation modeling and imputes data under an assumption that the data follow a multivariate normal distribution. The multivariate normal distributional assumption may not always hold, especially for multilevel hierarchical data with very small clusters. The aim of the present study is to demonstrate a method of multiply imputing missing values for data with a hierarchical or nested data structure using a well-known statistical software package. This approach is demonstrated using SUDAAN's HOTDECK procedure (SUDAAN Release 11, RTI International, Research Triangle Park, North Carolina) and then fit logistic models using the multiply imputed data.

Data

A population-based dataset collected from multiple sources was used. It included live birth records (2000-2006) from Florida, Massachusetts, and Michigan linked to the National Assisted Reproductive Technology (ART) Surveillance System (NASS) at the Centers for Disease Control and Prevention (CDC) (Centers for Disease Control and Prevention, 2014). The population of interest was infants conceived via ART. To eliminate the potential impact of subsequent treatments on maternal complications and pregnancy outcomes, only the first live born infant of the first live birth was included if a woman was identified as having more than one birth in the time period (Grigorescu, et al., 2014). Because the NASS data were reported by each fertility clinic in the United States, the data had a hierarchical structure and observations were nested in fertility clinics.

The main outcome of interest for our analysis was an Apgar score at five minutes, a binary variable corded as 0 (>=7) and 1 (<7). The Apgar score at five

minutes is the first test given to a newborn to quickly evaluate a newborn's physical condition with a score ranging from one to ten. Values of 7 and above are considered normal. The independent covariates in a logistic model were reason for ART (V_1), maternal age (V_2), race/ethnicity (V_3), education (V_4), adequacy of prenatal care (V_5), co-morbid conditions (V_6), delivery method (V_7), induction of labor (V_8), gestational age (V_9), newborn gender (V_{10}), and birth weight (V_{11}) (Grigorescu, et al., 2014).

Missing Value Imputation

SUDAAN was developed to analyze data from complex surveys; however SUDAAN is also able to analyze other hierarchical or nested data, or non-survey data. Data inspection showed that the amount of data missing for the outcome value was extremely small (<0.3%) so observations with missing outcome values were excluded, and imputed values only for observations with missing values for the covariates. SUDAAN's `HOTDECK` procedure was used to impute missing values of covariates, because 8.3% of the observations had a missing value for at least one covariate, resulting in a reduction of 67 cases. `HOTDECK` replaces missing values of one or more variables of a recipient using observed values from a "similar" respondent. Since our data were naturally clustered, i.e., the observations (infants) were clustered in fertility clinics, we restricted to obtaining the pool of respondents by clinic and replacing missing values of recipients in the same clinic. For each infant with missing values of the covariates (V_1, V_2, …, V_{11}), the HOTDECK procedure collected a set of similar infants from the same clinic (cluster) without missing covariates. From this set, randomly chosen infants were used to fill in the missing values of the covariates with replacement where each variable was filled separately. This process was repeated until all infants with missing values for covariates within the clinic were imputed. SUDAAN's `HOTDECK` procedure uses a weighted sequential hot deck method proposed by Cox (1980) and Iannacchione (1982) to perform imputation, the default method for `PROC HOTDECK`.

The SAS-callable SUDAAN was used with the following code for the `HOTDECK` procedure:

```
PROC HOTDECK DATA=DATA_INPUT SEED=3123845;
       IMPBY CLINIC;
       IMPID INFANT_ID;
       IMPVAR V1 V2 … V11/MULTIMP=5;
```

```
        WEIGHT _ONE_;
        IMPNAME V1="V1_IMP" V2="V2_IMP" … V11 = "V11_IMP";
        IDVAR APGAR;
        OUTPUT /IMPUTE=default FILENAME=OUTDATA REPLACE;
RUN;
```

In the PROC HOTDECK statement, DATA= specifies the input dataset (DATA_INPUT) which includes variables with missing values. The SEED= specifies an integer to generate a random number for the imputation. The cluster variable is specified on the IMPBY statement (CLINIC); data must be sorted by this cluster variable prior to running this procedure. Each observation clustered within the clinic is identified using the IMPID statement, in this case by the infant variable (INFANT_ID). The variables with missing values to be imputed (V_1, V_2, …, V_{11}) are listed in the IMPVAR statement. The option, MULTIMP=5, in the IMPVAR statement specifies that five imputed datasets are to be created. For the non-survey data, set the variable in the WEIGHT statement to be _ONE_, a default option in SUDAAN to indicate no weighting.

The IMPNAME statement assigns variable names for imputed variables (original variable name + IMP in our case). For each imputation, SUDAAN assigns a consecutive number after the imputed variable name (V1_IMP1 V2_IMP1 … V11_IMP1 in the first imputation, V1_IMP2 V2_IMP2 … V11_IMP2 in the second imputation, etc.). The IDVAR statement specifies that our outcome variable (APGAR), which was not imputed, should be included in the output dataset. The OUTPUT statement provides a dataset with all imputed variables, the cluster variable (specified by IMPBY), the imputation identification variable (specified by IMPID), and variables not imputed (specified by IDVAR). The option IMPUTE=default indicates that the output dataset will include all imputed variables ($11 \times 5 = 55$ imputed variables), the option FILENAME= specifies the name of the output dataset (OUTDATA), and the option REPLACE instructs SUDAAN to overwrite any existing dataset with the same name.

PROC MI in SAS (SAS v. 9.3, Cary, NC) was used to impute missing values in order to compare imputation results from PROC MI to those obtained from SUDAAN's PROC HOTDECK. The MI procedure is a parametric multiple imputation procedure that creates multiply imputed data sets using predicted values rather than observed values as HOTDECK to replace missing values. Due to some clinics having fewer than three observations (38.8% of total included clinics), PROC MI failed to provide any output for imputation. This demonstrates that the parametric imputation approach, such as sequential regression models, is limited in dealing

with very small clusters for multiple imputation. Because the `MI` procedure does not adequately perform imputation for the data, this method is not described in detail.

Statistical Analysis

Multiply imputed data was used. According to Rubin (1978), the multiple imputation estimator (denoted as $\hat{\theta}$) of parameter is the average of the estimators obtained from all K imputed datasets:

$$\bar{\theta}_K = \frac{1}{K}\sum_{i=1}^{K}\hat{\theta}_i, \tag{1}$$

The variance of $\bar{\theta}_K$ is the sum of the average within (imputed dataset)-imputation variance and the between (imputed datasets)-imputation variance. Because the population data was used, the finite population correction can be ignored, denoting the variance of the i^{th} imputed dataset as W_i, the average within-imputation variance is:

$$\bar{W}_K = \tfrac{1}{K}\sum_{i=1}^{K} W_i \tag{2}$$

and the between-imputation variance is:

$$B_K = \frac{1}{K-1}\sum_{i=1}^{K}\left(\hat{\theta}_i - \bar{\theta}\right)^2 \tag{3}$$

The overall variance of $\bar{\theta}_K$ is the sum of within-imputation variance and the between-imputation variance, with a bias correction for the finite number of multiply imputed data sets:

$$Var\left(\bar{\theta}_K\right) = \bar{W}_K + \tfrac{K+1}{K} B_K \tag{4}$$

The SAS-callable SUDAAN `RLOGISTIC` procedure was used to fit a random effects logistic regression model using imputed data. Collinearity was inspected between covariates using Zack's SAS Macro (n.d.) for the logistic model with the following `RLOGISTIC` procedure:

```
PROC RLOGIST DESIGN=WR DATA=IMPN1 MI_COUNT=5;
       NEST _ONE_ CLINIC;
       WEIGHT _ONE_;
       CLASS V1_IMP …;
       REFLEVEL V1_IMP=1 …;
       MODEL APGAR= V1_IMP V2_IMP … V11_IMP;
RUN;
```

In the PROC RLOGISTIC statement, set DESIGN = WR (sampling with replacement for population data, SUDAAN's default design). Using the output dataset from the imputation procedure (OUTDATA), we created 5 datasets (Sinharay, Stern, and Russell, 2001), one for each imputation, and each dataset included 14 variables, INFANT_ID, CLINIC, APGAR, V1_IMP, V2_IMP, …, V11_IMP for model fitting. Assign the names IMPN1, IMPN2, IMPN3, IMPN4 and IMPN5 to these datasets. The options DATA=IMPN1 and MI_COUNT=5 informs SUDAAN to use all five datasets (IMPN1, IMPN2, IMPN3, IMPN4, IMPN5) for pooling the estimates from the five logistic models. The statements NEST and WEIGHT are set for non-survey data that are nested within clinics (CLINIC). The CLASS statement is used to specify the categorical covariates and the REFLEVEL statement specifies the reference level for each categorical variable. Note with DESIGN=WR and the NEST and WEIGHT statements as listed, the variable CLINIC is modeled as a random effect.

Results

There were 335 cases with an Apgar score less than seven found in 16,833 infants in the data. The primary risk factor of interest was a three level (tubal obstruction only, ovulatory dysfunction only, and other reasons) variable of infertility diagnosis (reason for ART, V_1). The primary interest was in comparing women with ovulatory dysfunction only to women with tubal obstruction only, controlling for other covariates mentioned above. Using imputed data, all 335 cases were included in the adjusted model; however, only 268 cases and 15,430 infants could be used for the adjusted model derived from the original non-imputed data (20.0% less cases and 8.3% less infants). For our multivariable logistic model, the inspection of collinearity using Zack's SAS Macro showed that only one condition index is greater than 30, indicating no sign of multicollinearity between covariates.

The odds ratios, 95% confidence intervals (CI), and *P* values for the unadjusted and adjusted models for reason for ART are compiled in Table 1.

Comparing a diagnosis of only ovulatory dysfunction to only tubal factor, the unadjusted odds ratio (OR) using all 335 cases was 1.86 (95% CI: 1.31-2.63, P-value = 0.0005). Notice that the missing for V1 was negligible (comparing the imputed data adjusted odds ratio to the non-imputed data adjusted odds ratio) and no cases were deleted from the unadjusted analysis. Using the multiply imputed data, the adjusted odds ratio was 1.93 (95% CI: 1.31-2.84, P-value = 0.0009) and using the non-imputed data, the adjusted odds ratio was 1.73 (95% CI: 1.12-2.69, P = 0.015).

Table 1. Unadjusted odds ratio (OR) and adjusted odds ratio (aOR) for reasons for ART

Reason for ART	OR (95% CI*) P value	Imputed data aOR (95% CI*) P value	Non-Imputed data aOR (95% CI*) P value
Tubal Obstruction only	Ref	Ref	Ref
Ovulatory Dis-function only	1.86 (1.31-2.63) 0.0005	1.93 (1.31-2.84) 0.0009	1.73 (1.12-2.69) 0.015
Other reasons	1.20 (0.85-1.69) 0.297	1.35 (0.91-1.99) 0.134	1.27 (0.91-1.77) 0.152

**CI-Confidence interval*

Because there were a small number of infants with Apgar scores less than 7 (335/16,833), there was a concern that missing values of covariates would change the results of the adjusted model. This concern was addressed using the method of multiple imputation. Because the data were naturally clustered, consider the impact of such data structure in multiple imputation and modeling, which likely provides better statistical inferences than not addressing such impact on analysis. The SUDAAN HOTDECK procedure imputed missing values by incorporating covariate information in the imputation process. The merit of this approach is to use real (and hence realistic) values in imputation without strong parametric assumptions, and to provide good inferences for linear and non-linear statistics (Andridge & Little, 2010). However, this procedure has limitations, because it requires good matches of respondents to recipients based only on available covariate information and finding good matches is more likely in large clinics. Moreover, repeating the HOTDECK with the same respondent pool but randomly sorting data is an arguable imputation procedure. To determine the impact of this method on the results, we also conducted the analysis using the traditional complete observations method. In this study, the results were similar, meaning

multiple imputation may not be necessary. However, the conclusion does not exclude the possibility that results may vary across applications.

The data had a hierarchical or nested data structure with observations (infants) clustered within fertility clinics. The impact of this data structure was addressed in the multiple imputation and statistical analysis using the SUDAAN software package. The example provided could be applied to other datasets with hierarchical or nested structures where missing values of variables are a concern.

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

References

Andridge, R. R. and Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review, 78*(1), 40–64. doi: 10.1111/j.1751-5823.2010.00103.x

Centers for Disease Control and Prevention, American Society for Reproductive Medicine, Society for Assisted Reproductive Technology. (2014). *2012 Assisted Reproductive Technology Success Rate*. Atlanta, GA: Centers for Disease Control and Prevention.

Cox, B. (1980). The weighted sequential hot deck imputation procedure. In *American Statistical Association Proceedings* Survey Research Methods Section, pp. 721–726. Alexandria, VA: American Statistical Association. Retrieved from http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1980_152.pdf

Grigorescu, V., Zhang, Y., Kissin, D., et al. (2014). Maternal characteristics and pregnancy outcomes after assisted reproductive technology (ART) by infertility diagnosis: ovulatory dysfunction (OD) versus tubal obstruction (TO). *Fertility and Sterility*, 101(4), 1019–1025. doi: 10.1016/j.fertnstert.2013.12.030

Iannacchione, V. (February, 1982). *Weighted sequential hot deck imputation macros*. Paper presented at the Seventh Annual SAS User's Group International Conference, San Francisco, CA.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis, 9*(2), 137–163. doi: 10.1093/oxfordjournals.pan.a004868

Rotnitzky, A. and Wypij, D. (1994). A note on the bias of estimators with missing data. *Biometrics*, *50*(4), 1163–1170. doi: 10.2307/2533454

Rubin, D. B. (1978). Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse. In *American Statistical Association Proceedings* Survey Research Methods Section, pp. 20–34. Alexandria, VA: American Statistical Association. Retrieved from http://ww2.amstat.org/sections/SRMS/Proceedings/papers/1978_004.pdf

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. NY: John Wiley and Sons. doi: 10.1002/9780470316696

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489. doi: 10.1080/01621459.1996.10476908

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, *8*(1), 3–15. doi: 10.1177/096228029900800102

Sinharay, S., Stern, H., and Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*(4), 317–329. doi: 10.1037/1082-989x.6.4.317

Zack, M. (n.d.) SAS Macro [Computer software]. Retrieved from http://schick.tripod.com/collin.sas