MWSUG 2017 - Paper AA06

Multiple Imputation of Family Income Data in the 2015 Behavioral Risk Factor Surveillance System

Jia Li, National Institute for Occupational Safety and Health (NIOSH), Cincinnati, OH
Aaron L. Sussell, National Institute for Occupational Safety and Health (NIOSH), Cincinnati, OH

ABSTRACT

Multiple imputation methods are increasingly used to handle missing data in statistical analyses of observational studies to reduce bias and improve precision. SAS/STAT® PROC MI can be used to impute continuous or categorical variables with a monotone or arbitrary missing pattern. This study used the fully conditional specification (FCS) method to impute the family income variable in the 2015 Behavioral Risk Factor Surveillance System (BRFSS) data. BRFSS is a state-based health survey that collects data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. In this paper, the study population was restricted to currently employed respondents (age>=18) from the 25 states that collected industry and occupation information. Of the total 87,483 respondents, 11% were missing income information.

To impute the missing income data, all variables in the survey that are correlated with either income or missingness of income (N=28) were selected as covariates. BRFSS sample design variables that represent stratification and unequal sampling probabilities were also included in the imputation model to improve validity. The FCS method was chosen due to an arbitrary missing pattern and mixed data types among income and all covariates. Logistic regression and discriminant function options were used for imputing binary and ordinal/nominal variables respectively. Results show a significantly different distribution in imputed income values compared to the observed values, suggesting that using the traditional complete case analysis approach to analyze BRFSS income data may lead to biased results.

INTRODUCTION

Missing data is a common problem in statistics. The traditional complete case analysis of data, which excludes cases with missing data for any variable in a proposed model, can result in biased estimates and loss of statistical power if the data have a substantial amount of missing values (>5%). Multiple imputation (MI), which is based on the assumption that data are missing at random (MAR), has become one of the most used methods to deal with missing data. In contrast to several single imputation methods which include hot-deck, cold-deck, mean substitution, and regression imputation, MI has the ability to account for the additional uncertainty introduced by imputation.

In a large observational study, data can be missing on a mixed set of continuous, nominal, ordinal, and count variables, and the missing pattern is often arbitrary. The Fully Conditional Specification (FCS) method (Van Buuren, 2007) can be used for imputation of this type of data. FCS imputations are generated sequentially by specifying a multivariate model for each variable given the other covariates. It specifies a set of conditional densities, one for each incomplete covariate. Starting from an initial imputation, FCS draws imputations by iterating over the conditional densities.

Generally, for imputation of a particular variable, the model should include variables that are correlated with the imputed variable, as well as variables that are associated with the missingness of the imputed variable (Little & Rubin, 2002). In a complex survey setting, the sample design variables should also be included in the imputation model for the validity of the resultant multiply imputed dataset (Rubin, 1987).

DATA AND ANALYSES

The Behavioral Risk Factor Surveillance System (BRFSS) is a state-based cross-sectional telephone survey. The BRFSS questionnaire is comprised of a core section and several optional modules. BRFSS sample design involves stratification and unequal sampling probability. Stratification is done by state, substate geographic region, and type of phone (landline vs cell phone) interviews. There are a total of 692 strata in BRFSS 2015 data.

In our current study the population of interest is employed adults (age>=18) from 25 states that conducted the optional industry and occupation survey module. Of the 100,700 respondents meeting these criteria, 13,217 (13%) were excluded due to missing or uncodable industry and occupation information. The remaining data have 9,930 (11%) respondents with missing income.

In this study, income was collapsed into 4 categories: less than $25K, $25-50K, $50-75k, $75K+. A binary indicator of missingness of income was also created. BRFSS 2015 core section data have a total of 110 variables. Including a large amount of unimportant predictors in imputation models can lead to possible lost precision (Little & Rubin, 2002). In this study, only variables correlated with either income or missingness of income were selected as candidate covariates for the imputation model. As some of these potential covariates may be correlated among themselves, logistic regressions were conducted with stepwise model selection methods to remove redundant variables.

To capture the complex sample design features in the imputation model, a combination of state and landline/cell phone sample identifiers was created. The survey weights were categorized into five equal-sized groups. These two sample design variables were included in the imputation model.

RESULTS

Pairwise associations between each variable and income, as well as associations between each variable and missingness of income were computed with PROC FREQ. Thirty five variables with Cramer's V greater than 0.1 were selected as potential candidates as covariates for imputation of income. After stepwise logistic regression model selection, 28 covariates were kept (Table 1). Table 2 lists covariates among the 28 that are associated with missingness of income.

Table 1. BRFSS 2015 employed adults -- Covariates included in the imputation model for income

Covariate Name	Covariate Label
_AGE_G	Imputed age in six groups
_PRACE1_R	Race
_HISPANC	Ethnicity
_EDUCAG	Education
MARITAL_R	Marital status
OCC_NHIS	Occupation NHIS Simple 2-digit Recode of Census Codes
IND_NHIS	Industry NHIS Simple 2-digit Recode of Census Codes
STATE_PHONE	State by phone type
_FINALWT_C	Sample weight categorized into 5 groups
QSTLANG_R	Language used for survey
RENTHOM1	Own or rent home
CPDEMO1	Have a cell phone for personal use
INTERNET	Internet use in the past 30 days
GENHLTH	General health
HLTHPLN1	Have any health care coverage
PERSDOC2	Multiple health care professionals
MEDCOST	Could not see doctor because of cost
CHOLCHK	How long since cholesterol checked
TOLDHI2	Ever told blood cholesterol high
BLIND	Blind or difficulty seeing
DECIDE	Difficulty concentrating or remembering
DIFFWALK	Difficulty walking or climbing stairs
ARTHDIS2	Arthritis affects work
FLUSHOT6	Adult flu shot/spray past 12 months
_SMOKER3	Smoking status
DRNKANY5	Drank any alcoholic beverages in past 30 days
_VEGLT1	Consume vegetables 1 or more times per day
_PACAT1	Physical activity categories

Table 2. BRFSS 2015 employed adults -- Distribution of missingness of income by selected covariates

Covariate	% missing income
Imputed age in six groups	
Age 18 to 24	24.1
Age 25 to 34	9.9
Age 35 to 44	8.0
Age 45 to 54	10.1
Age 55 to 64	11.5
Age 65 or older	14.6
Own or rent home	
Own	10.1
Rent	11.7
Other arrangement	28.9
Education	
Did not graduate High School	15.6
Graduated High School	14.2
Attended College or Technical School	11.1
Graduated from College or Technical School	9.4

The code used for imputation of the 4-category income variable using the FCS method is shown below:

```
%LET VARLIST = STATE_PHONE _FINALWT_C _AGE_G IND_NHIS OCC_NHIS QSTLANG_R
CPDEMO1 GENHLTH _EDUCAG MEDCOST HLTHPLN1 RENTHOM1 PERSDOC2 MARITAL_R _HISPANC
INTERNET _PRACE1_R ARTHDIS2 BLIND DIFFWALK DECIDE TOLDHI2 _SMOKER3 CHOLCHK
DRNKANY5 FLUSHOT6 _VEGLT1 _PACAT1 INCOMEGRP4;
PROC MI DATA = TEMP NIMPUTE = 10 OUT = BRFSS.IMPUTED SEED = 1234;
      CLASS &VARLIST;
      VAR &VARLIST;
      FCS NBITER=20
      LOGISTIC(QSTLANG_R CPDEMO1 MEDCOST HLTHPLN1 MARITAL_R _HISPANC INTERNET
            ARTHDIS2 BLIND DIFFWALK DECIDE DRNKANY5 FLUSHOT6 _VEGLT1
            /LIKELIHOOD=AUGMENT DETAILS)
      DISCRIM(GENHLTH _EDUCAG RENTHOM1 PERSDOC2 _MRACE1_R TOLDHI2 _SMOKER3
            CHOLCHK _PACAT1 INCOMEGRP4
            /CLASSEFFECT=INCLUDE DETAILS);
      ODS EXCLUDE MISSPATTERN;
RUN;
```

The VAR statement lists all variables to be included in the model. Those variables with missing values in the list were imputed in the process even though only income was the one of interest. The FCS method imputes variables sequentially in the order specified in the VAR statement. Although changing the variable order will result in minor changes to the imputed data, the impact of the ordering of variables can be reduced by increasing the number of imputed datasets. In this study, the variables were listed in the order of increasing percentage of missing to improve computational efficiency. The NIMPUTE=10 option in the PROC MI statement specifies ten imputed datasets to be created. The NBITER=20 option in the FCS statement specifies 20 burn-in iterations before imputation. Logistic regression was used to impute binary variables, and the discriminant function method was chosen for ordinal and nominal variables. The first five covariates in the VAR statement have no missing data and were omitted from the LOGISTIC/DISCRIMINANT options of the FCS statement. The LIKELIHOOD=AUGMENT suboption in the LOGISTIC option requests maximum likelihood estimates based on augmented data, which solves from the problem of a quasi-complete separation data pattern.

The distribution of imputed income is shown in table 3 along with the distribution of observed income.

Table 3. BRFSS 2015 employed adults -- Comparison of observed and imputed income distribution

Family income	Observed %	Imputed %	Overall %
<$25,000	17.3	28.9	18.7
$25,000-$50,000	23.4	23.8	23.4
$50,000-$75,000	17.6	14.7	17.2
>$75,000	41.7	32.6	40.6

CONCLUSION

PROC MI is an imputation procedure that creates multiple imputed datasets for an incomplete multivariate dataset. Missing values are replaced with a set of plausible values that represent a random sample and thus represent the uncertainty about the missing value.

For the 2015 BRFSS data, the distribution of imputed income generated in this study is significantly different from the distribution of the observed income. Ignoring the missing income data while using complete case analysis may lead to potential bias.

The FCS method is a useful tool for creating imputations in large data sets with complex data structures such as arbitrary missing patterns and mixed types of variables.

MI is based on the assumption that data are missing at random (MAR). MAR is not a testable assumption. Future studies will focus on sensitivity analysis for an assumption of missing not at random (MNAR).

REFERENCES

Rubin DB. 1987. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons.

Little RJ, Rubin DB. 2002. Statistical analysis of missing data. New York: John Wiley & Sons.

Van Buuren, S. 2007. Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification. Statistical Methods in Medical Research 16:219–242.

Liu Y, De A. 2015. Multiple Imputation by Fully Conditional Specification for Dealing with Missing Data in a Large Epidemiologic Study. Int J Stat Med Res 4(3): 287–295.

CONTACT INFORMATION

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

Jia Li
NIOSH
QZL0@CDC.GOV