Imputation for Longitudinal Study of Effectiveness of an Anti-Smoking Campaign

Qiao Ma[1], Edward Mulrow[1], Josiane Bechara[1], Zachary H. Seeskin[1]
Morgane Bennett[2], Jennifer Cantrell[2], Elizabeth Hair[2], Donna Vallone[2]

[1]NORC at the University of Chicago, 55 E. Monroe Street, Chicago, IL 60603
[2]900 G Street, NW, Fourth Floor, Washington, DC 20001

Abstract
The Truth Initiative Longitudinal Cohort Study is designed to evaluate the impact of a television and digital campaign on youths' smoking-related knowledge, attitudes and beliefs, perceived social norms, and behaviors over time. The study administers surveys to participants over six waves between 2014 and 2017 and uses multivariate statistical models to evaluate the effectiveness of the media campaign. The survey is subject to nonresponse, which can bias estimates for the evaluation. We describe and examine different methods of imputing missing data in the context of a longitudinal study. Hot deck and model-based approaches are compared for both their performance and practicality. Examining income, the variable with the highest item nonresponse rate, we find that using either hot deck or model-based estimation helps correct for nonresponse bias in estimates from complete case analysis, and we demonstrate how multiple imputation can help account for the uncertainty in estimates due to imputation.

Key Words: Missing Data, Hot Deck Imputation, Multiple Imputation, Evaluation, Nonresponse Bias

1. Introduction

Imputation is a commonly used method to handle item nonresponse for surveys. When a case has item missing data, imputation fills in values for the missing data instead of excluding cases with missing data from the analysis. The objective of imputation is not to get the best possible predictions of the missing values, but to fill in missing values in such a way as to allow for inference about population parameters (Little & Rubin 2002).

One commonly used imputation method is hot deck imputation (Rubin 1987). This method does not rely on model fitting for the variable to be imputed. Thus, hot deck imputation is potentially less sensitive to model misspecification than are imputation methods based on a parametric model, such as regression imputation (Andridge & Little 2010).

This paper discusses imputation for the Truth Initiative Longitudinal Cohort study (TLC). The study is designed to evaluate the impact of a television and digital campaign on youths' smoking-related knowledge, beliefs, perceived social norms, and behaviors over time. Participants are followed over six waves between 2014 and 2017. Prior studies found nonresponse significantly related to smoking behavior, which can affect estimates of the final evaluation model. Thus, imputation is important for mitigating the effects of nonresponse bias.

Hot deck and model-based imputation used for TLC imputation are described, and a comparison is conducted to evaluate the techniques for imputing income, the variable with the highest item nonresponse rate. We also discuss how multiple imputation helps account for the uncertainty in estimates due to missing data.

2. Description of Imputation Approach

In survey methodology, missing data can arise due to nonresponse. There are two types of nonresponse: unit nonresponse and item nonresponse. Unit nonresponse is when a sample member provides no information (Rubin 1987). For example, the individual is not available when selected for a survey. Item nonresponse, on the other hand, is when a unit offers incomplete information, only partially completing a survey. This study focuses on the latter of the two types of nonresponse.

Many tools have been proposed for dealing with missing data problems. Among the most widely used tools for imputation procedures include mean imputation, hot deck imputation, and regression imputation. Our approach primary used two techniques: weighted sequential hot deck and model-based imputation.

2.1 Weighted Sequential Hot Deck Procedure (WSHD)

Weighted sequential hot deck procedure (WSHD) uses survey data from respondents as donors to provide imputed values for records with missing values. Imputation cells are defined usually based on a cross-classification of covariates. Then, missing values are replaced with donor values selected within each imputation cell. WSHD makes use of the survey weight so that the weighted distribution of imputations reflects the weighted distribution of the available data.

For TLC, WSHD was first applied to impute demographic variables. We used age, gender and education to define cross-classification donor cells in order to conduct the imputation for the other demographic variables with missing values, such as race, parental race, and parental education.

After missing demographic items were imputed, we proceeded with completing non-demographic, non-income items. To determine the variables defining the WSHD cells, we ran random forest models using the R package *randomForest*. Parental education and region consistently had the largest measures of variable importance, so these variables were used to form the hot deck donor cells for most variables remaining to be imputed.

2.2 Model-Based Imputation

As income was the variable with by the highest item nonresponse rate (26.8% in Wave 1), we used a model-based approach for its imputation in order to capture a richer set of covariates. Specifically, we used predictive mean matching involves the following steps:

- On an initial iteration, missing values are filled in.
- A linear regression model is fit to the data at each iteration, calculating estimates of the regression parameters and their posterior distributions.
- Then, values of the regression parameters are randomly drawn (selected) from these posterior distributions.
- Afterwards, predicted values are calculated for each observation using the randomly drawn regression parameters.
- Each missing value has an imputation selected from among the 5 observed values with the nearest predictions.

- The process iterates 20 times until convergence is achieved.

Further, we provided multiple imputations of income to allow a user to account in for the uncertainty in any estimates due to item nonresponse for income. Multiple imputation fills in the missing data m times to generate m complete datasets. Rubin (1987) describes that each complete dataset should be analyzed by standard procedures. Then, if desired, the results of the m analyses can be combined to estimate parameters and their standard errors, reflecting the uncertainty in estimates due to imputation.

3. Comparison of Imputation Approaches for Income

3.1 Hot Deck and Predictive Mean Matching

The estimates of the income distribution comparing using the two imputation methods to complete case analysis are shown in Table 1. Both hot deck and predictive mean matching correct for the nonresponse bias of complete case analysis. Complete case analysis underestimates the proportion with income less than $19,999 and overestimates the proportion with income over $100,000.

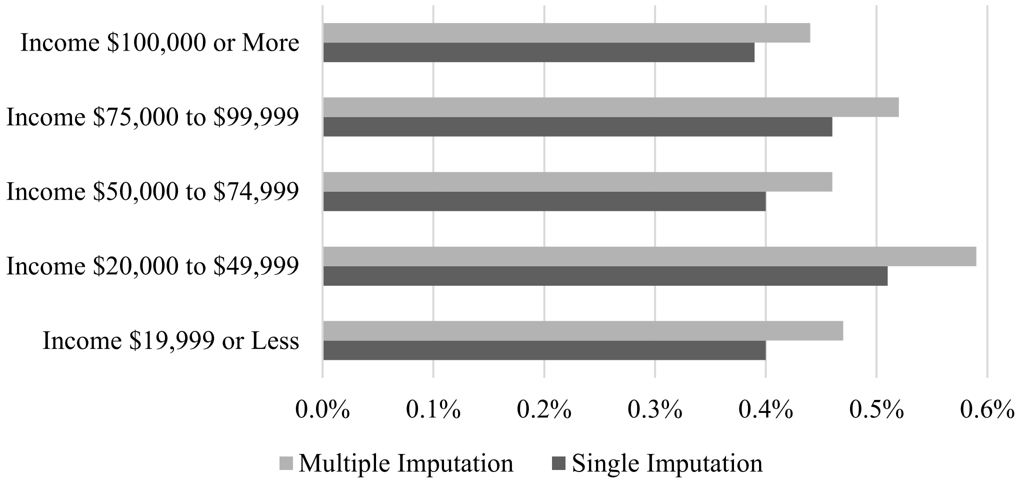
Table 1: Estimates of Percentage of Households in Different Income Groups by Method

Income Group	Complete Case	Hot Deck Imputation	Model-Based Imputation
Income $19,999 or Less	11.8%	12.9%	13.3%
Income $20,000 to $49,999	25.8%	25.7%	25.9%
Income $50,000 to $74,999	16.4%	16.7%	16.6%
Income $75,000 to $99,999	27.1%	26.8%	26.6%
Income $100,000 or More	19.0%	18.0%	17.6%

3.2 Comparing Single and Multiple Imputation

Figure 1 compares the standard errors between single imputation and multiple imputation by income group, using the model-based imputation approach. As seen, single imputation can understate the uncertainty of estimates, as the standard errors do not reflect the uncertainty due to missing data. Single imputation underestimates these standard errors by about 10 to 15 percent.

Figure 1: Comparisons of Standard Errors for Single Imputation vs. Multiple Imputation



4. Discussion

Imputation is a useful approach to address item nonresponse in surveys and mitigate the effects of nonresponse bias on estimates. We describe two approaches implemented for the Truth Initiative Longitudinal Cohort study. Hot deck imputation is straightforward to conduct and less sensitive to model misspecification. When imputing a variable with high item nonresponse, a model-based approach can incorporate a richer set of covariates. Further, model-based approaches are compatible with multiple imputation. As shown, accounting for the uncertainty in estimates due to imputation can be critical, and multiple imputation is one approach for doing so.

References

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, **78**, 40-64, 2010.

Ault, K. (2013). Imputation Methods for Surveys: A Demonstration of the Impute Procedure in SUDAAN. *In Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 389–398

Cox, B. G. (1980). The weighted sequential hot deck imputation procedure. *In Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 721–726.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed., New York: Wiley

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.