An Imputation Model Database and Its Relevance to Analysis

Peter Frechtel[1], Glynis Ewing[2], Kristen Gulledge[2]
Susan Edwards[2], Jonaki Bose[4]
[1]RTI International, One Metro Center, 701 13th Street NW, Suite 750, Washington, DC 20005
[2]RTI International, 3040 Cornwallis Road, Research Triangle Park, NC 27709
[3]Substance Abuse and Mental Health Services Administration, 1 Choke Cherry Road, Rockville, MD 20857

Abstract

The National Survey on Drug Use and Health (NSDUH) is sponsored by the Substance Abuse and Mental Health Services Administration and provides national, state and substate data on substance use and mental health in the civilian, noninstitutionalized population age 12 and older. The NSDUH is a continuous survey, with approximately 67,500 interviews completed annually. As part of the NSDUH imputation procedures, over 400 regression models are fit each year. These models are used to match each item nonrespondent with a "neighborhood" of similar item respondents in order to identify a donor. The response variables in these models are variables of primary interest to analysts. After the procedures are complete for each year, an imputation model database is populated which stores covariate-level information such as the p-values associated with the regression coefficients. This database can be used both by staff working on the NSDUH imputation and by staff analyzing the NSDUH data. This paper illustrates how such a database can be used not only by those conducting the imputation, but also by those making decisions during the analysis of NSDUH data.

Key Words: imputation; model-based imputation, predictive mean neighborhoods

1. Introduction

The National Survey on Drug Use and Health (NSDUH) is sponsored by the Substance Abuse and Mental Health Services Administration and provides national, state and substate data on substance use and mental health in the civilian, noninstitutionalized population age 12 and older. The NSDUH is a continuous survey, with approximately 67,500 interviews completed annually.

As part of the regular imputation procedures, the NSDUH imputation team fits over 400 regression models each year. These models are used to obtain predicted values for variables that undergo imputation. The predicted values are then used to match each item nonrespondent with a "neighborhood" of similar item respondents. One of the members of the neighborhood is selected as the donor, and the missing value for the item nonrespondent is replaced by the nonmissing value of the donor. The method is described in detail in Laufenberg et al. (2014).

The imputation team has recently developed a "model database," which stores covariate-level information from each model. The database includes results of hypothesis tests that the regression coefficients are nonzero. This is useful for several reasons. First, these results can be used to reduce the starting list of covariates a priori, which would be an easy way to speed up the procedures should that be desired. (During imputation processing, significant time is spent reducing the covariate lists in order to achieve convergence in the models.) Second, these results can be used to develop starting lists for variables which will undergo imputation in the future. Third, the results can be used to inform decisions on the order in which the variables are processed and the grouping of variables into "imputation sets." But fourth, and perhaps most importantly, the results can be used in analysis. Many of the variables that undergo imputation are of primary interest to analysts, including drug measures such as recency, frequency, and age at first use. Although there are some caveats, the imputation model results can be mined for analysis ideas, or can be used to see quickly whether a hypothesized relationship between variables is likely to exist.

To expand on the last point, there is generally a close relationship between imputation models and data analysis, especially statistical modeling. It is easiest to see this by considering the model-based approach to sampling (Royall 1970). Under this approach, the outcome variable is simply the realization of a model. It is not a fixed value attached to a unit as it is in design-based sampling, where the only random process is the drawing of the sample. Ideally the model used in imputation would be the same as the theoretical model used to generate values for the outcome variable. The theoretical model used to generate values for the outcome variable is exactly what the typical data analyst is interested in.

Jerry Reiter has said that to his way of thinking, everything is a missing data problem (personal communication, March 9, 2012). When you sample from a finite population, the values for the unsampled units are realizations of a model, and so are the values for nonrespondents among the sampled units. Any analysis to him involves using (imputation) models to fill in missing data for some survey variables among the unit respondents, and all survey variables for both the unit nonrespondents and the unsampled units.

Another demonstration of the connection between imputation models and analysis is the generation of synthetic data sets (e.g., Raghunathan et al., 2003). Designed to protect confidentiality, these data sets are generated using complex models, and the complex models are built using the real (confidential) data. The synthetic data is a lot like imputed data. The process by which the models are fit to the real data may be similar to the process an analyst might undertake to discern relationships between the survey variables.

The purpose of this document is to describe the imputation model database and to stimulate a discussion on how it might be used in analysis.

2. Description of the Model Database

The Imputation Model Database (IMD) is a SAS data set that is created after all imputation procedures have been completed for a given survey year. The rows in the

IMD associated with a single regression model are presented in Table 1. The columns are as follows:

- Drop Flag: 0 if the variable was included in the final model, and 1 if the variable was dropped. The IMD includes all covariates that were in the starting list for the model.
- Wald F: the statistic associated with the hypothesis test that the regression coefficient is nonzero.
- p-value: the p-value associated with that hypothesis test.

Table 1: The rows of the covariate level IMD for cocaine 30-day frequency of use model, respondents aged 12-17 (n=29, respondents with past month use)

Variable	*Drop Flag*	*Wald F*	*P-value*
X_1	0	0.3038	0.5816
X_2	1		
X_3	1		
X_4	0	0.1596	0.6896
X_5	0	1.4377	0.2308
X_6	1		
X_7	1		
X_8	1		
X_9	1		
X_{10}	0	0.7697	0.3805
X_{11}	0	0.1699	0.6803
X_{12}	0	1.2959	0.2744

For the model shown in Table 1, 12 covariates were in the starting list, and 6 were in the final list. There were some variables in the final model that did not seem to help much with prediction conditional on these other variables. During imputation processing, prediction is the primary goal and parsimony is not required. Thus, these variables were kept in the imputation model.

3. How the Model Database Might Be Used in Analysis

An example of how the IMD may be used in analysis is below. Table 2 presents the IMD entries for the gender covariate for all lifetime drug use models for respondents aged 12-17. An analyst might draw the preliminary conclusion that, controlling for other factors, gender is highly correlated with most of the lifetime use indicators, but not so much for crack, marijuana, and pipes.

Table 2: IMD entries for the gender covariate in lifetime use drug models, respondents aged 12-17

	Number of Covariates in Model		*Gender Covariate*		
Drug	*Starting List*	*Final List*	*P-value*	*Rank of p-value among all drug models*	*Rank of p-value among final covariates*

Smokeless Tobacco	22	22	0.0000	1	2
Cigars	24	24	0.0000	2	3
Alcohol	26	26	0.0000	3	8
Tranquilizers	31	29	0.0000	4	5
Cocaine	34	21	0.0002	5	2
Stimulants	32	29	0.0007	6	6
Sedatives	33	26	0.0086	7	4
Inhalants	27	27	0.0098	8	9
Pain Relievers	30	29	0.0131	9	11
Hallucinogens	29	29	0.0568	10	12
Heroin	36	11	0.1774	11	4
Pipes	25	25	0.2661	12	17
Marijuana	28	28	0.6266	13	24
Crack	34	10	0.9638	14	10

4. Limitations on the Relevance of the IMD to Analysis

Some of the methods used to build the imputation models make them less useful for analysis. These models certainly should not be treated as a final step in analysis. It is more reasonable to use them as a preliminary step. Some limitations of using the IMD for analysis are:

- Little attention is given to parsimony when fitting the models. We tend to drop only as many covariates as we need to in order to get the model to converge. As a result, there are probably a lot of noise parameters in the majority of the models, making it difficult to tell whether the statistically significant ones are really helpful.
- Sometimes the best covariates are dropped from the models because they trigger certain warning messages in SUDAAN. The warning messages are produced when a cross-classification of the outcome variable and the covariate has empty or nearly empty cells. It is the best covariates that tend to produce empty or nearly empty cells. An analyst would never drop these covariates.
- Even if a covariate was dropped in the imputation model, it may still be a good predictor for the dependent variable. If other covariates in the model were removed, this covariate may become significant.
- Many of the models have such a large sample size that most of the covariates are significant. Covariates with limited predictive power may be statistically significant due only to the large sample size.
- The IMD currently does not store the estimates for the regression coefficients (i.e., the "betas"), or even their signs (positive or negative). The estimates themselves may not be useful for analysis because they may be difficult to interpret. For linear models, the response variables in the linear models sometimes undergo transformations that make interpretation difficult. For logistic models, interpretation of the estimates is tricky for those not experienced with odds ratios and the like. Still, the sign may be useful. One option to increase the

utility of this database would be to add the estimates and their standard errors to the IMD for future NSDUH years.

- So far, the IMD has only been populated for the 2012 and 2013 NSDUHs. It would be difficult and time-consuming to populate it for earlier years. Limited conclusions may be drawn from two years of data.

5. Next Steps

The IMD was built for use by the NSDUH imputation team. The data sets used to populate it are created naturally while the imputation procedures are run, and the process of populating it each year is straightforward. It seems a shame that all these models are being fit every year that may be of interest to analysts, but no analysts ever see them.

We also hope to demystify model-based imputation procedures for the benefit of analysts. Terms like "Predictive Mean Neighborhoods" and "model-based imputation method" sound intimidating, but there's nothing intimidating about the modeling process. It's hardly different from what statisticians and data analysts learn in school.

Acknowledgements

The authors would like to thank Jim Chromy, Dan Liao, Phil Kott, Karol Krotki, Valerie Hoffman, and Jeremy Aldworth for their helpful comments and suggestions on this document. The authors would also like to thank Andrew Moore, Cynthia Augustine, Joey Morris, and Jamie Ridenhour for their helpful contributions to the accompanying presentation.

References

Laufenberg, J., Kroutil, L., Frechtel, P., Carpenter, L., Edwards, S., Ewing, G., Gulledge, K., Handley, W., Martin, P., Moore, A., & Scott, V. (2014). Editing and imputation report. In *2012 National Survey on Drug Use and Health: Methodological resource book* (Sections 10/11, prepared for the Substance Abuse and Mental Health Services Administration, Contract No. HHSS283201000003C, Deliverable No. 41, RTI/0212800.001.107.006.007). Research Triangle Park, NC: RTI International.

Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. (2003). "Multiple Imputation for Statistical Disclosure Limitation," Journal of Official Statistics, 19, 1-16.

Royall, R.M. (1970). "On Finite Population Sampling Theory Under Certain Linear Regression Models," Biometrika, 57, 377-387.