

ROLES FOR BAYESIAN TECHNIQUES IN SURVEY SAMPLING

Trivellore E. Raghunathan¹ and Donald B. Rubin²

ABSTRACT

The standard randomization-based approach to survey sampling is dominant in current practice and is ideally suited to the task of evaluating how statistical procedures will operate in long run practice. The Bayesian approach, in contrast, is ideally suited to the task of providing guidance on how to create procedures in difficult situations. These approaches are complementary, not competitive. A striking example of this complementary nature of Bayesian and frequentist modes of thinking in survey sampling is the use of multiple imputation to address nonresponse. By following the Bayesian paradigm, procedures are created that have frequentist operating characteristics that are superior to previous ad hoc methods. This conclusion is dramatically illustrated by the results of a project involving missing data in the Consumer Expenditure Survey from the U.S. Bureau of Labor Statistics.

KEY WORDS: Confidence coverage, Consumer Expenditure Survey, Ignorable missing data mechanism, Multiple Imputation, Sampling properties, Simulation.

RÉSUMÉ

L'approche fréquentiste utilisée dans l'échantillonnage est dominante en pratique et convient idéalement à l'évaluation du comportement à long terme d'une procédure. Contrairement à l'approche fréquentiste, l'approche bayésienne est particulièrement adaptée à la création de nouvelles procédures pour des cas difficiles. Ces approches sont complémentaires et ne sont pas en compétition. Un exemple frappant de la complémentarité des modes de penser bayésien et fréquentiste dans l'échantillonnage est l'utilisation de l'imputation multiple pour la non-réponse. En suivant le cadre bayésien, on crée des procédures qui possèdent des propriétés fréquentistes qui sont supérieures aux méthodes ad hoc précédentes. Cette conclusion est illustrée à l'aide des résultats d'un projet comprenant des données manquantes dans l'enquête sur les dépenses des consommateurs du U.S Bureau of Labor Statistics.

MOTS CLÉS: Taux de recouvrement, Consumer Expenditure Survey, mécanisme de non-réponse ignorable, imputation multiple, propriétés d'échantillonnage, simulation.

1. INTRODUCTION: FREQUENTIST EVALUATIONS AND BAYESIANLY DERIVED PROCEDURES

The standard randomization-based approach to survey sampling, due to Neyman (1934) and developed by Cochran, Madow, and others, is well established. It basically evaluates how a procedure operates in long run practice, and requires that (a) estimates are approximately unbiased for their estimands, (b) hypothesis tests have at most their nominal rejection rates when null hypotheses are true, and (c) confidence intervals cover their estimands, with at least their

nominal levels. For standard survey practice, such evaluations of proposed estimates, test procedures and confidence intervals are very relevant, especially when done with appropriate conditioning. The limitation of the approach, however, is that it provides essentially no useful guidance for creating procedures in complicated settings. As a result, it has encouraged a continuing proliferation of overly Baroque and unprincipled procedures, which are inferior with respect to their randomization-based operating characteristics.

In contrast to the randomization-based paradigm, the Bayesian paradigm is ideally suited for creating

¹ Institute for Social Research, University of Michigan, Ann Arbor, MI 48106.

² Department of Statistics, Harvard University, Cambridge, MA 02138

procedures based on principled considerations relating to predicting unobserved values. If basic inferential principles are adhered to, all evidence we know supports the contention that following the Bayesian paradigm is a direct path for creating procedures with excellent randomization-based operating characteristics.

An analogy with solving "word problems" is helpful. Such problems, as given for example to students in grammar school, can often appear very difficult especially when there are two or more unknowns (e.g. John's age, his brother's age, and their mother's age). Once basic algebra is learned, however, the "calculus" of algebra makes the derivation of the answer relatively straightforward. Certainly, the algebraic approach is more straightforward than ad hoc trial and error methods that evaluate various proposed solutions to see if they work. The Bayesian approach is the calculus for deriving statistical procedures and the frequentist approach corresponds to the evaluation of the proposed answer. Because of the greater uncertainty of statistical inferences relative to answers to logical word problems, frequency evaluations may be more critically important than checking the answers in word problems, although few of us should trust our algebraic or modeling skills so much that we eschew evaluating proposed answers!

There is a dramatic and well-studied example in survey practice of the complementary nature of Bayesian and frequentist methods. This is multiple imputation for nonresponse, where in complex settings, it remains the only feasible approach. Despite recent criticisms by some (e.g. Fay (1996)), multiple imputation is not only generally valid, but it is also highly efficient in all realistic examples where it has been studied. A recent frequentist evaluation of multiple imputation using data from the Consumer Expenditure Survey at the US Bureau of Labor Statistics will illustrate the reason for this conclusion. Other recent works on multiple imputation where procedures derived using the Bayesian approach have been shown to have desirable frequency properties include Raghunathan and Grizzle (1995) for dealing with split questionnaire designs (or matrix sampling of items) and Raghunathan and Siscovick (1996) for dealing with the analysis of case-control data with missing covariate values.

2. MULTIPLE IMPUTATION EXAMPLE

To exemplify the issues discussed in Section 1, we describe the results from a simulation study that was conducted as a part of a Task Order under the contract

given to the University of Michigan (UofM) by the Bureau of Labor Statistics (BLS). This study was set up so that missing data were created by one side (BLS) but multiply imputed by another side (the authors) who did not know how the missing data were created. Our multiple imputations, despite having been created by a Bayesian model, were evaluated strictly by their frequentist operating characteristics.

In addition, although many now accept the value of multiple imputation, some at BLS preferred fixing parameter estimates at their maximum likelihood estimates rather than drawing values from their posterior distributions as required by the general theory (e.g., Rubin (1987)). Moreover, some economists are concerned with the use of all variables to predict the missing values, the exogeneity/endogeneity problem. As a result, different versions of multiple imputation were studied.

The basic situation involved the Consumer Expenditure Survey where all data were considered observed except for income and race, which were sometimes missing. The other variables included expenditures of various types and background variables. The specific objective of this project was to evaluate the following two issues:

1. Whether or not the expenditure variables should be used as predictors when the missing income data are multiply imputed, and
2. whether the fully Bayesian approach for multiply imputing the missing income data has better inferential properties from a frequentist perspective than the approach currently being proposed by the BLS staff, which does not draw all parameters from their joint posterior distribution.

This report is preliminary but we believe compelling. A subsequent report by all participants will provide details.

3. SIMULATION SETUP

For the simulation study, BLS staff created a pseudo-population by accumulating complete-cases (that is, the subjects with no missing data) over several years of the survey; this is analogous to the initial setup in an NCHS project (Ezzati-Rice et al. (1995)). Next BLS staff drew two hundred simple random samples, each of approximate size 500 from this population. For each of the 200 samples, for approximately 30% of the

individuals the income data were deleted using an ignorable missing data mechanism as defined in Rubin (1976), so that all the variables used in creating the missing data were included in the data sets provided for creating imputations. All these steps were undertaken by the BLS staff alone, and we did not know how the population was created or the nature of the mechanism that was used to delete income and race data from each of the 200 samples. The BLS staff then provided the 200 data sets with partially missing income and race data along with complete data on 55 covariates, including the expenditure variables, that would be used in the subsequent substantive analyses of the multiply imputed data sets. We (the multiple imputers) did not know (and still do not know) which of these covariates were used in the missing data mechanism.

4. CREATION OF THE MULTIPLY IMPUTED DATA SETS

Each of the 200 data sets was then multiply imputed using the following four approaches; only the first one is fully justified by multiple imputation theory.

Method 1 is a fully Bayesian approach where an explicit model for the joint distribution of the variables with missing values conditional on the fully observed covariates was first postulated. Next, using a noninformative prior for the unknown parameters in this model, the posterior predictive distribution of the missing values was constructed. Finally, the missing values were imputed using draws from this posterior predictive distribution.

Specifically, suppose that I denotes the log-income variable, E denotes the log-expenditure variables, and X denote the covariates except race/ethnicity, which is coded as White ($R = 1$), Hispanic ($R = 2$), Black ($R = 3$) and Other ($R = 4$). In these 200 data sets, there were no missing values in E and X . The missing values in I and R were multiply imputed as follows.

Using the notation $[A|B]$ for the conditional distribution of A given B , we modeled the joint distribution of I and R as

$$[I|R, E, X, \phi][R|E, X, \theta],$$

where ϕ and θ are the unknown parameters. The imputations for those with missing R were created by drawing values from the posterior predictive distribution $[R|E, X]$ and then conditioning on the imputed values of R , missing I values were drawn from the

conditional posterior predictive distribution $[I|R, E, X]$. The distribution $[R|E, X, \theta]$ was modeled using a polytomous regression model with

$$\log \left[\frac{\Pr(R = j|E, X)}{\Pr(R = 4|E, X)} \right] = \alpha_j + \beta_j^t E + \gamma_j^t X,$$

where $j = 1, 2, 3$, and $\theta = (\alpha_j, \beta_j^t, \gamma_j^t; j = 1, 2, 3)$ denotes the parameters in this model. Suppose that $\hat{\theta}$ and V , respectively, denote the maximum likelihood estimate of θ and its asymptotic covariance matrix. Our strategy was to draw θ from its asymptotic multivariate normal distribution $N(\hat{\theta}, V)$ and then draw R from the multinomial logit model conditional on the drawn value of θ .

The following describes the essential steps for imputing R :

1. Define $\theta_* = \hat{\theta} + V^{1/2}z$ where z is a vector of random normal deviates of dimension $rows(V)$ and V is the Cholesky decomposition of V (that is, $V(V)^t = V$). Reshape the vector θ_* into a matrix B_* of dimension $3 \times columns(U)$ where $U = [1, X, E]$.

2. Let U_{miss} denote the rows of U with missing R and let

$$P_{j*} = \exp [U_{miss} B_{j*}] / \left[1 + \sum_j \exp [U_{miss} B_{j*}] \right],$$

where B_{j*} is the j^{th} column of B_* where $j = 1, 2, 3$ and $P_{4*} = 1 - \sum_j P_{j*}$.

3. Let $S_0 = 0$, $S_j = \sum_i^j P_{i*}$, $j = 1, 2, 3$ and $S_4 = 1$ be the cumulative sums of the cell probabilities.

4. To impute missing values, generate a vector of uniform random numbers u of dimension $rows(U_{miss})$ and take j as the imputed category if the corresponding components satisfy the inequality $S_{j-1} \leq u \leq S_j$.

Next, to specify $[I|R, E, X]$, we considered a linear regression model,

$$I = \alpha + \beta^t E + \gamma^t X + \delta^t R + \epsilon,$$

where R is now coded as 3 dummy variables, $\phi = (\alpha, \beta^t, \gamma^t, \delta^t)$ are the unknown

regression coefficients, and ϵ are the error terms assumed to be independent normal random variables with mean zero and common variance σ^2 . Let $U = [1, E, X, R]$ denote the design matrix of predictors. Suppose that $\hat{\phi} = (U'U)^{-1}U'I$ is the estimated regression coefficient and $s^2 = (I - U\hat{\phi})'(I - U\hat{\phi})/df$ is the estimated residual variance, where $df = (\text{rows}(I) - \text{columns}(U))$ is the degrees of freedom. The following steps then provide draws from the posterior predictive distribution of missing values of the variable I .

1. Generate a chi-square random deviate c with df degrees of freedom and compute $s_*^2 = df \times s^2/c$.
2. Generate a vector $z = (z_1, z_2, \dots, z_p)'$, (of dimension $p = \text{rows}(\phi)$) of random normal deviates and define $\phi_* = \hat{\phi} + s_* \times V^{1/2} z$ where V is the Cholesky decomposition of $(U'U)^{-1}$.
3. Let U_{miss} denote the U -matrix for those with missing I values; then the imputed set of values is $I_* = U_{\text{miss}} \phi_* + s_* z$, where z is another independent vector (of dimension $\text{rows}(U_{\text{miss}})$) of random normal deviates.

To create multiple imputations, we repeated the two steps of drawing first from $[R | E, X]$ and then drawing from $[I | R, E, X]$ just described. A total of 5 imputations for all the missing values were created for each of the 200 data sets.

Method 2: This method followed exactly the steps for **Method 1** as outlined above except that the expenditure variables E were *not* included as predictors.

Methods 3 & 4: When imputing R and I under Method 1, the estimated regression coefficient $\hat{\theta}$ and $\hat{\phi}$ were perturbed across the multiple imputations using their estimated covariance matrices to reflect the uncertainty in the parameter estimates. We created two additional sets of multiply imputed data sets, paralleling Method 1 and Method 2 that do not perturb the estimated covariance matrices, but fix them at their maximum likelihood estimates. These are not proper imputations as defined by Rubin (1987) but has been proposed by BLS in their earlier work

5. FREQUENTIST EVALUATION OF MULTIPLE IMPUTATION INFERENCES

We considered two possible complete-data substantive analyses of interest to compare the frequentist operating characteristics of these four approaches. The first substantive analysis involved a multiple linear regression model,

$$I = \alpha_0 + \alpha_1 \times E + \sum_j^p \alpha_j Z_j,$$

where $Z_{pj} = 1, 2, \dots, p$ are the covariates representing AGE, EDUCATION and RACE. The second substantive analysis involved regressing E on I ,

$$E = \beta_0 + \beta_1 \times I + \sum_j^p \beta_j Z_j.$$

The primary parameters of interest are α_j and β_j .

The BLS staff know the true value of α_j and β_j by running these regression models on the entire population. We constructed 200 nominal 90% and 95% for α_j and β_j for each method of imputation using the standard multiple imputation formulas (repeated imputation inferences, Rubin, 1987). We then computed the proportion of intervals that contained the appropriate true value. Table 1 tabulates the results for both the parameters of interest.

Table 1: The exact levels of the confidence intervals using the four approaches for multiple imputation; Only Method 1 is based on Bayesian Theory.

Imputation Method	Parameter of Interest			
	α_1		β_1	
	90%	95%	90%	95%
Method 1	91.5	96.0	90.5	95.5
Method 2	82.5	88.0	81.5	87.5
Method 3	89.0	92.0	88.0	91.5
Method 4	80.0	86.5	79.5	85.5

Clearly, from the results in Table 1, the fully Bayesian approach results in well-calibrated confidence intervals though they are slightly conservative. The intervals from Method 3 are slightly anti-conservative because of the underestimation in the multiple imputation standard errors due to ignoring uncertainty in parameter estimation. In contrast, the intervals under Methods 2 and 4, which do not include the expenditure variables in

the imputation process, result in intervals that are extremely anti-conservative. Not including the expenditure variables in the imputation model is tantamount to assuming a zero partial correlation coefficient between the I and E variables conditional on X and R , which results in substantial bias towards the null in the point estimates from each completed data set. Such problems are inherent in many multivariate analysis when the imputations are not conditional on all the observed values.

This simulation, which will be described in more detail elsewhere, clearly demonstrates the complementary roles in sample survey work of using Bayesian methods to create procedures and frequentist methods to evaluate them: Following the Bayesian paradigm can lead to procedures with superior frequentist performance.

ACKNOWLEDGEMENTS

The authors thank the BLS staff, especially Geoffrey Paulin, for creating and providing the data sets with missing values used in this project. This work was performed for the Task Order # M-11 under the Contract # J-9-J-5-0025 from BLS to the University of Michigan.

REFERENCES

Ezzati-Rice, T., Johnson, W., Khare, M., Little, R.J.A., Rubin, D.B., and Schafer, J.L. (1995). A simulation study to evaluate the performance of model-based multiple imputation in NCHS health examination surveys. *Bureau of the Census Eleventh Annual Research Conference*, 257-288.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of American Statistical Association*, 91, 490-498.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of Royal Statistical Society*, 97, 558-606.

Raghunathan, T.E. and Grizzle, J.E. (1995). A split questionnaire survey design. *Journal of American Statistical Association*, 90, 54-63.

Raghunathan, T.E. and Siscovick, D.S. (1996). Multiple

imputation analysis of case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensives. *Applied Statistics*, 45, 335-352.

Rubin, D.B. (1976). Inference and missing data (with discussion). *Biometrika*, 63, 581-592.

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

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of American Statistical Association*, 91, 434-489.