Weights, Double Protection, and Multiple Imputation

Phillip S. Kott[1] and Ralph E. Folsom[2]
[1]RTI International, 6110 Executive Blvd., Suite 902, Rockville, MD 20852
[2]RTI International, 3040 Cornwallis Rd., Research Triangle Park, NC 27709

Abstract
In its purest form, multiple imputation is a technique that compensates for item nonresponse using prediction modeling. Although developed in a Bayesian framework, its advocates claim the technique has good "frequentist" properties. With weighted survey data, however, this is generally true only when the item missingness is completely at random. We present a way to conduct a weighted multiple imputation under which resulting estimates are doubly protected from nonresponse bias; that is to say, if either the assumed prediction model or the response (propensity) model is correct, the resulting estimator is nearly unbiased in some sense. Unfortunately, the multiple-imputation-variance estimator will itself be nearly unbiased only when both models hold. Unlike multiple imputation, available imputation and variance-estimation techniques requiring only one of the two models to be true generally focus on a single survey item at a time.

Key Words: Bootstrap, Item-response model, Prediction model**,** Survey weight, **N**onresponse bias

1. Introduction

Rubin (1987, 1996) introduced the technique of multiple imputation (more correctly labeled "repeated imputation") for handling item nonresponse in complex surveys and measuring its impact of mean squared error. Although developed in a Bayesian framework, Rubin claimed that this technique had good "frequentist properties" under certain conditions.

Many took that to mean the multiple-imputation variance-estimator had good properties if inferences under an assumed *prediction model* (relating the survey variable to covariates) were replaced by inference under the probability-sampling mechanism and an assumed *response model* (governing which units respond to the item in question and which don't). Looking at a few simple special cases of survey-weighted estimates, Kott (1995) showed this to be a misunderstanding. Item nonresponse had to be completely at random (independent of all covariates) for the multiple-imputation variance-estimator to be nearly unbiased. Moreover, it appeared that the prediction model had to hold as well. Kim *et al.* (2006) put Kott's observations in a more rigorous and general framework.

Given a complex survey data set and a fitted item-response model, we will propose a new method for conducting a multiple imputation and estimating its variance. We begin with a brief description of multiple imputation. A discussion of prediction modeling for single imputation follows. In it, we introduce the notion of a quasi-random response model and show how to conduct prediction-mean imputation in such a way that the resulting survey estimate is doubly protected from nonresponse bias under mild conditions. We then present a simple bootstrapping method for conducting multiple imputation that can also result in an estimator doubly protected from nonresponse bias. This is followed by an

investigation of the multiple-imputation variance estimator in this context, which, unlike the multiple imputation itself, effectively requires both the prediction and response models to be correct for the variance estimator to be unbiased. We end with some concluding remarks.

2. A Brief Outline of Multiple Imputation

To simplify the exposition, we focus on the imputation of a single item subject to nonresponse and the impact of that nonresponse on the estimation of a population mean for that item. Note that a proportion is a special case of a mean.

Let y denotes the item and k the element. In multiple imputation, a missing y_k is imputed T times (T is often chosen to be 5 in practice) and then the average of those imputations is computed. If each of the T imputations is denoted $\tilde{y}_{kt}$ (t = 1,..., T), then the final imputed value for y_k is

$$\tilde{y}_k = \tfrac{1}{T}\sum_{t=1}^{T}\tilde{y}_{kt}. \tag{1}$$

It is more common to think of multiple imputation as the combination of T completed data sets (samples). One such data set D_t consists of y_k for each $k \in R$ and $\tilde{y}_{kt}$ for each $k \in M$, where R denotes the subset of the sample with valid responding y-values, and M. the subset without valid responding y-values.

Suppose that in the absence of item nonresponse the estimator for the population y-mean, P, has the form:

$$p_F = \frac{\sum_S d_k y_k}{\sum_S d_k}, \tag{2}$$

where S denotes the sample and d_k an element sampling weight (after, perhaps, adjusting for *unit* nonresponse and calibrating for sample balance). The multiple-imputation estimator would then be

$$p_{MI}^{T} = \frac{\sum_R d_k y_k + \sum_M d_k \tilde{y}_k}{\sum_S d_k}, \tag{3}$$

where R denotes the set of r sample units providing item y-values, and M the set of m item nonrespondents.

Another way to express this estimator is as

$$p_{MI}^{T} = \frac{1}{T}\sum_{t=1}^{T} p_{MI}^{(t)}, \text{ where } p_{MI}^{(t)} = \frac{\sum_R d_k y_k + \sum_M d_k \tilde{y}_{kt}}{\sum_S d_k}. \tag{4}$$

Observe that $p_{MI}^{(t)}$ is computed by treating all y-members of the completely data set D_t as a full sample.

The multiple-imputation variance estimator has the form:

$$v(p_{MI}^{T}) = B(1+\tfrac{1}{T}) + W, \tag{5}$$

where $W = \left(\sum^{T} \text{var}(p_{MI}^{(t)})\right)\Big/T$ is the so-called "within variance." It estimates the variance of p_F, the estimator of the mean had there been no nonresponse. Each var($p_{MI}^{(t)}$) is calculated as if the singularly-imputed $\tilde{y}_{kt}$ were equal to the real missing y_k.

The B is equation (5) is

$$B = \frac{1}{T-1}\sum_{t=1}^{T}(p_{MI}^{(t)} - p_{MI}^{T})^2, \tag{6}$$

the so-called "between variance." It estimates the contribution to variance caused by imputation in the theoretical construct p_{MI}^{∞}, which is what would result if multiple imputation had been conducted an infinite number of times. This contribution to variance is formally E[($p_{MI}^{\infty} - p_F$)^2].

3. Prediction Modeling, Survey Weights, and Single Imputation

Suppose every y_k, whether item respondent or nonrespondent, is assumed to fit a prediction model of the form:

$$y_k = \mu_k + \varepsilon_k, \tag{7}$$

where the ε_k are random variables with mean zero given μ_k, and each $\mu_k = \mu(\mathbf{x}_k\boldsymbol{\beta})$ is a function of a known row vector of covariates $\mathbf{x}_k$ including 1 (or the equivalent), while $\boldsymbol{\beta}$ is an unknown column vector of parameters.

Two common examples of the function $\mu(\mathbf{x}_k\boldsymbol{\beta})$ are $\mu_k = \mathbf{x}_k\boldsymbol{\beta}$, which can be reasonable when y_k is continuous, and $\mu_k = (1 + \exp(-\mathbf{x}_k\boldsymbol{\beta}))^{-1}$, which can be sensible when y_k is binary (0/1). We effectively limit our treatments here to these examples.

Often $\mathbf{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kG})$ is a vector of group-membership indicators ($x_{kg} = 1$ when k is in group g, 0 otherwise).When these groups are mutually exclusive, equation (7) is a group-mean model.

If the goal is to estimate the population mean of the y_j, then the ideal imputation for a missing y_k would be the *predictive mean*, μ_k, assuming the prediction model in equation (7) is correct. But μ_k is unknown because $\boldsymbol{\beta}$ is unknown. If $\mathbf{b}$ were a consistent estimator for $\boldsymbol{\beta}$, then the *estimated predictive mean*, $\hat{\mu}_k = \mu(\mathbf{x}_k\mathbf{b})$, would be a nearly (i.e., asymptotically) unbiased estimator for μ_k.

An intriguing method for determining $\mathbf{b}$ is given below. It incorporates both the sampling weights, d_k, and estimates of the element probabilities of item response, ρ_k. We will

assume these ρ_k have been computed based on a quasi-random response model independent of the prediction model in equation (1).

We propose finding a **b** that satisfies

$$\sum_R d_k \frac{1-\rho_k}{\rho_k}\left(y_k - \mu(\mathbf{x}_k\mathbf{b})\right)\mathbf{x}_k = \sum_S d_k \frac{1-\rho_k}{\rho_k} r_k \left(y_k - \mu(\mathbf{x}_k\mathbf{b})\right)\mathbf{x}_k = \mathbf{0}, \tag{8}$$

where r_k is a random variable equal to 1 when k is an item respondent and 0 otherwise.

The choice of the implicit weight $w_k = d_k(1-\rho_k)/\rho_k$ in equation (8) assures us that

$$\sum_R d_k \frac{1-\rho_k}{\rho_k} y_k = \sum_R d_k \frac{1-\rho_k}{\rho_k} \hat{\mu}_k, \tag{9}$$

where $\hat{\mu}_k = \mu(\mathbf{x}_k\mathbf{b})$ since effectively 1 is a component of $\mathbf{x}_k$.

To see why this weighting may be useful will take a bit of work. Let us assume the quasi-random model generating the element item-response probabilities is correct and ignore the finite-sample distinction between the ρ_k and the response probabilities they estimate (this distinction commonly goes away asymptotically). Let us also ignore the finite-sample distinction between **b** and the solution, $\mathbf{b}^0$, to

$$\sum_S d_k \left(1-\rho_k\right)\left(y_k - \mu(\mathbf{x}_k\mathbf{b}^0)\right)\mathbf{x}_k = \mathbf{0}, \tag{10}$$

where the left-hand side of equation (10) is the expectation under the response mechanism. Finally, assume that a solution to equation (8) exist *whether or not the prediction model*, $E(y_k) = \mu(\mathbf{x}_k\boldsymbol{\beta})$, holds.

With this in mind, taking the response expectation of both sides of equation (9) reveals that

$$\sum_S d_k (1-\rho_k) y_k \approx \sum_S d_k (1-\rho_k)\hat{\mu}_k. \tag{11}$$

The above is an approximate equality because of the finite-sample distinctions we are ignoring. This means that if the response model is correct, then imputing for missing y_k with $\hat{\mu}_k$ produces an unbiased estimator in some sense *even if the prediction model in equation (7) does not hold.*

Similarly, if the item-response model does not hold, but the prediction model does, then this imputation approach is unbiased in a prediction-model sense. Consequently, using equation (8) to estimate **b** results in imputations that have been called "doubly protected" (or "doubly robust") against item nonresponse. See, for example, Bang and Robins (2005).

4. A Bootstrapped Version of Weighted Multiple Imputation

As pointed out earlier, with multiple imputation a missing y_k is imputed T times. Rather than deriving T estimates of μ_k, a good multiple-imputation method provides T predictions of $y_k = \mu_k + \varepsilon_k$. We propose doing that using the following steps:

1. For each set of m imputations in D_t, draw a simple random sample *with* replacement of size r from the r elements in R. This is called the t^{th} bootstrap respondent sample and is denoted by R_t.

2. Estimate $\mathbf{b}_t$ for each t by replacing R is equation (8) with R_t (using ρ_k).

3. Compute $\hat{\mu}_{kt} = \mu(\mathbf{x}_k\mathbf{b}_t)$ for each missing y_k.

4a. If y is binary (0/1), then independently set each y_{kt} to 1 with probability $\hat{\mu}_{kt}$ and to 0 otherwise.

4b. If y is continuous, then compute $e_{jt} = y_j - \mu(\mathbf{x}_j\mathbf{b}_t)$ for every unit in R_t. Set $y_{kt} = \mu(\mathbf{x}_j\mathbf{b}_t) + e_{(k)t}$, where each of the m residuals $e_{(k)t}$ is selected from among the e_{jt} in R_t with probability proportional to $w_j = d_j(1-\rho_j)/\rho_j$ either with or without replacement.

There are appealing alternatives to 4b (for example, ones that force y_{kt} to be nonnegative when the y-values are all nonnegative), but the formulation the serves our immediate purposes. Observe that **b** satisfying equation (8) assures that the expected value of $e_{(k)t}$ under the selection mechanism in 4b is 0.

Under a group-mean model with constant w_j within groups, the bootstrap described above using Step 4b is very similar to the approximate Bayesian bootstrap (Rubin and Schenker 1986) except that in that methodology a separate bootstrap sample is drawn in every group.

The bootstrap selection mechanisms in Step 1 and both versions of 4 do not rely on questionable modeling assumptions. They are fully under our control. For simplicity, we will assume the same for the original sampling mechanism associated with the d_k even though, in reality, it may incorporate unit-nonresponse adjustments.

Suppose the full-item-response estimator is equation (2) is nearly unbiased under the sampling mechanism. We show in the next section that the multiple-imputation estimator in equation (4) is nearly unbiased in some sense if either the prediction model in equation (7) or the item-response model generating the ρ_j is correct. When only the prediction model is correct, the estimator is nearly unbiased under the combination of the original-sampling and bootstrap selection mechanisms and the prediction model. When only the item-response model is correct, the estimator is nearly unbiased under the combination of the original-sampling, bootstrap, and item-response selection mechanisms.

5. The Multiple-Imputation Variance Estimator

5.1 The Decomposition

Let p_{MI}^{∞} be this idealized multiple-imputation estimator based on an infinite number of sets of imputations. Observe that

$$p_{MI}^{\infty} - P = (p_M^{\infty} - p_F) + (p_F - P),$$

so that

$$(p_{MI}^{\infty} - P)^2 = (p_{MI}^{\infty} - p_F)^2 + (p_F - P)^2 + 2(p_{MI}^{\infty} - p_F)(p_F - P). \tag{12}$$

Multiple-imputation attempts to estimate the expected value of the quantity on the left with equation (5). It does this by estimating the expectation of the first two terms on the right. These estimates are *B* and *W* in equation (5). The expectation of the third term in equation (12) is assumed to be asymptotically zero. It can be, as we shall see, when the expectation is taken over both the prediction and response models.

It is not hard to see that the arithmetic mean of the infinite imputations for a missing y_k described last section would be its estimated predictive mean, μ_k. If $\mu(z)$ is a smooth function, then

$$\hat{\mu}_k \approx \mu_k + \hat{\mu}_k' \mathbf{x}_k \left(\sum_R w_j \hat{\mu}_j' \mathbf{x}_j^T \mathbf{x}_j \right)^{-1} \sum_R w_j \mathbf{x}_j^T \varepsilon_j,$$

where $\hat{\mu}_k'$ is the first derivative of $\mu(\mathbf{x}_k\mathbf{b})$. Consequently,

$$\begin{aligned} p_{MI}^{\infty} - p_F &= \frac{\sum_M d_k (\hat{\mu}_k - y_k)}{\sum_S d_k} \\ &\approx \frac{\sum_M d_k \left(\left[\mu_k + \hat{\mu}_k' \mathbf{x}_k \left(\sum_R w_j \hat{\mu}_j' \mathbf{x}_j^T \mathbf{x}_j \right)^{-1} \sum_R w_j \mathbf{x}_j^T \varepsilon_j \right] - \left[\mu_k + \varepsilon_k \right] \right)}{\sum_S d_k} \\ &= \frac{\sum_M d_k \hat{\mu}_k' \mathbf{x}_k \left(\sum_R w_j \hat{\mu}_j' \mathbf{x}_j^T \mathbf{x}_j \right)^{-1} \sum_R w_j \mathbf{x}_j^T \varepsilon_j - \sum_M d_k \varepsilon_k}{\sum_S d_k}, \end{aligned} \tag{13}$$

while

$$p_F - P = \left(p + \frac{\sum_S d_k \varepsilon_k}{\sum_S d_k} \right) - P = \frac{\sum_S d_k \varepsilon_k}{\sum_S d_k} + \left(\frac{\sum_S d_k \mu_k}{\sum_S d_k} - P \right). \tag{14}$$

Since p_F is an unbiased estimator for *P*, for p_{MI}^{∞} to be doubly protected from nonresponse, we need the right-hand side of equation (13) to have mean zero under the prediction model, which it clearly does, and expectation near zero under the item-response model, which we show soon. Note that $p_{MI}^{(t)}$ is an unbiased estimator for p_{MI}^{∞} under the bootstrap selection mechanisms, so p_{MI}^{T} is doubly protected when p_{MI}^{∞} is.

5.2 The Last Term of Equation (12)

Let us look first at the last term on the right-hand side of equation (12). This term needs to be small for the multiple-imputation variance estimator in equation (5) to be useful. Under the prediction model in equation (7), with uncorrelated ε_k each with variance σ_k^2:

$$E_\varepsilon\left[(p_{MI}^{\infty} - p_F)(p_F - P)\right] = \frac{\sum_M d_k \hat{\mu}_k \,'\mathbf{x}_k \left(\sum_R w_j \hat{\mu}_j \,'\mathbf{x}_j{}^T \mathbf{x}_j\right)^{-1} \sum_R d_j w_j \mathbf{x}_j{}^T \sigma_j^2 - \sum_M d_k^2 \sigma_k^2}{\left(\sum_S d_k\right)^2}. \qquad (15)$$

The right-hand side of equation (15) is not zero. To make it nearly so, we can assume the item-response model. We then have the asymptotic equalities:

$$\sum_R w_k C_k \approx \sum_S d_k \tfrac{1-\rho_k}{\rho_k} r_k C_k \approx \sum_S d_k \tfrac{1-\rho_k}{\rho_k} C_k \rho_k = \sum_S d_k C_k (1-\rho_k), \text{ and}$$

$$\sum_M d_k C_k = \sum_S d_k C_k (1-r_k) \approx \sum_S d_k C_k (1-\rho_k).$$

From this, we see that the numerator of the right-hand side of equation (15) is asymptotically equal to

$$\sum_S d_k \hat{\mu}_k \,'\mathbf{x}_k (1-\rho_k)\left(\sum_S d_j \hat{\mu}_j \,'\mathbf{x}_j{}^T \mathbf{x}_j (1-\rho_j)\right)^{-1} \sum_S d_j^2 \mathbf{x}_j{}^T \sigma_j^2 (1-\rho_j) - \sum_S d_k^2 \sigma_k^2 (1-\rho_k).$$

Recall that either a component of $\mathbf{x}_k$ is 1 or the equivalent. This means there is a vector $\mathbf{g}$ such that $\mathbf{x}_k \mathbf{g} = \mathbf{g}^T \mathbf{x}_k{}^T = 1$ for all k. Accordingly, we can rewrite $\sum_S d_k \hat{\mu}_k \,'\mathbf{x}_k (1-\rho_k)$ as $\sum_S d_k \hat{\mu}_k \,'\mathbf{g}^T \mathbf{x}_k{}^T \mathbf{x}_k (1-\rho_k)$. After some work, the expression above collapses to 0.

Observe that when there is no explicitly postulated response model, the ρ_k are implicitly assumed to be identical. Put another way, nonresponse is assumed to be completely at random.

5.3 A Domain Mean

Even when both the prediction and response models are correct, the value of $E\left[(p_{MI} - p_F)(p_F - P)\right]$ may not be nearly zero for a domain mean. This can happen because the estimation of the prediction-model parameter $\boldsymbol{\beta}$ uses respondent information from outside of the domain.

The equivalent of equation (15) in this context can be shown to be

$$E\left[(p_{MI}^{\infty} - p_F)(p_F - P)\right] = \frac{\sum_M d_k z_k \hat{\mu}_k \mathbf{'x}_k \left(\sum_R w_j \hat{\mu}_j \mathbf{'x}_j{}^T \mathbf{x}_j\right)^{-1} \sum_R d_j z_j w_j \mathbf{x}_j{}^T \sigma_j^2 - \sum_M d_k^2 z_k \sigma_k^2}{\left(\sum_S d_k z_k\right)^2},$$

where $z_k = 1$ when k is in the domain of interest, and 0 otherwise. Notice that z_j is missing from $(\sum_R w_j \hat{\mu}_j \mathbf{'x}_j{}^T \mathbf{x}_j)^{-1}$. This will tend to make the third component and the mean squared error of p_M^{∞} negative. As a result, the multiple-imputation mean-squared-error will tend to be biased upward. This bias will be ignorably small when z_j is a component of $\mathbf{x}_j$ and the domain has its own prediction model (e.g., each component of the **x**-vector that has a nonzero value for an element *in* the domain has zero values for every element *outside* the domain).

5.4 Estimating the Between Variance

Observe that arguments analogous to those used above to show that the right-hand side of equation (15) is approximately zero under the item-response model (except, sometimes, for a domain mean) can be applied to the right-hand side of equation (13). This is the last piece we needed to establish the double protection of the multiple-imputation estimator, p_{MI}^T.

Unfortunately, the square of that expression (which is the sum of the variances of $(\sum_S d_k)^{-1} \sum_M d_k \hat{\mu}_k \mathbf{'x}_k (\sum_R w_j \hat{\mu}_j \mathbf{'x}_j{}^T \mathbf{x}_j)^{-1} \sum_R w_j \mathbf{x}_j{}^T \varepsilon_j$ and $(\sum_S d_k)^{-1} \sum_M d_k \varepsilon_k)$ is assured of having an expected value nearly equal to the expected value of B in equation (6) only when the original-sampling and bootstrap selection mechanisms are combined with the prediction model. Given that our bootstrap samples elements from the original respondent sample, *the* ε_k *may need to be uncorrelated*. *Moreover, for continuous variables, the variance of the* ε_k *may need to be equal* (to capture the variance of $(\sum_S d_k)^{-1} \sum_M d_k \varepsilon_k$ correctly).

5.5 Estimating the Within Variance

Let us now turn to the combined original-sampling/prediction variance of the without-item-imputation estimator, p_F. *Assuming the* ε_k *are uncorrelated*, we can see from equation (14) that this combined variance is the sum of the variance of

$$p_{F,\mu} = \frac{\sum_S d_k \mu_k}{\sum_S d_k}$$

under the original sampling mechanism and $\sum_S d_k^2 \sigma_k^2 / (\sum_S d_k)^2$. When the y-variable is binary, σ_k^2 is a function or μ_k, and our bootstrap imputation routine estimates μ_k in a nearly unbiased fashion assuming the prediction model holds. In contrast to this, when the y-variable is continuous, W in equation (5) will be a nearly unbiased estimator of this sum *when the* σ_k^2 *are constant* but not necessarily otherwise.

6. Concluding Remarks

Although the bootstrap method introduced here to multiply impute for item nonresponse was essentially new, the results in the paper do no depend on it. Double protection for item nonresponse results only from the way the parameter $\boldsymbol{\beta}$ was estimated in equation (8). The limitations of the multiple-imputation variance estimator in equation (5) are inherent in the way the variance is decomposed in equation (12). They have little to do with how the components are estimated.

If the only goal of imputation were to estimate population means (or totals), then one should impute for a missing y_k with the estimated predictive mean, $\hat{\mu}_k = \mu(\mathbf{x}_k\mathbf{b})$, where $\mathbf{b}$ satisfies equation (8). The resulting estimator would be doubly protected against nonresponse. Moreover, if the without-item-imputation estimator, p_F, were unbiased under the original sampling mechanism (including, perhaps, weighting for unit nonresponse), then it would not be hard to derive a jackknife that estimates the combined sampling/ prediction-model variance in a nearly unbiased fashion. In addition, the ε_k can be hetero-scedastic and correlated within primary sampling units (but not across them).

Although the double-protection against nonresponse bias from predictive-mean imputation as described above does not extend to an estimated domain mean when the membership indicator for the domain is not a covariate of the model, both the domain mean and its estimated jackknife variance can be nearly unbiased under the combination of the original sampling mechanism and the prediction model.

Often one has additional goals in mind other than estimating means when imputing for item nonresponse in survey data. In particular, estimating the distribution of variables and the relationship between variables can also be of interest.

Imputing for a missing continuous y_k initially with $y_{kt} = \mu(\mathbf{x}_j\mathbf{b}_t) + e_{(k)t}$ is more conducive to the estimation of the distribution of the population y-values than predictive-mean imputation. By taking the mean of T such imputations, multiple-imputation almost perfectly let's the user "have his cake and eat it too" (perfection obtains when $T = \infty$).

A price the cake enthusiast apparently has to pay is to assume that the ε_k are uncorrelated. If not, their correlation structure needs to be modeled and the impact of that structure worked into the imputation process. How to do such a thing is beyond the scope of this analysis.

Assuming the ε_k are uncorrelated, one may want to employ a different or more refined method for choosing the $e_{(k)t}$ than offered in Step 4b. Under a heteroscedastic error model, there are several things that can be done. When the σ_k^2 are known (or estimated) up to a constant multiple, one can replace Step 4b with

4b*. If y is continuous, then compute $e_{jt} = y_j - \mu(\mathbf{x}_j\mathbf{b}_t)$ for every unit in R_t. Set $y_{kt} = \mu(\mathbf{x}_j\mathbf{b}_t) + \sigma_k e_{(k)t}/\sigma_{(k)}$ where $e_{(k)t}$ is selected from among the e_{jt} in the R_t with probabilities proportional to $w_j = \sigma_j d_j(1-\rho_j)/\rho_j$ either with or without replacement.

An appealing alternative that assumes only that the variance is a function of $\mu(\mathbf{x}_j\mathbf{b})$ is to sort both R_t and M by their respective $\mu(\mathbf{x}_j\mathbf{b}_t)$-values, choose the donors using systematic probability sampling and assign them systematically to the missing y_k.

In a similar vein, one can sort the entire sample by their $\mu(\mathbf{x}_j\mathbf{b})$-values, use this sort to divide the sample into G roughly-equal groups with limited variability of $\mu(\mathbf{x}_j\mathbf{b})$-values within each group. One then remodels the y_k using the group-mean model, draws bootstrap samples independently within each group, and uses these samples to create the donors as in Section 3.

A little will be lost when the prediction model in equation (7) is correct, but if the functional form of $\mu(.)$ is other than what has been specified, this approach implicitly computes (and bootstraps) a locally linear approximation of the true functional form.

Throughout this paper, we have ignored the issue of how to estimate the ρ_k. That was because any estimation method for the ρ_k that produces nearly unbiased estimates under the response model would be sufficient for our purposes. For example, we employed the response model to show that the multiple-imputation variance estimator could be nearly unbiased when measuring the mean squared error under a combination of the sampling mechanism and prediction model. Since this mean squared error did not include a response-model component, estimating the response-model parameters has no large-sample impact on its size.

References

Bang H. and Robbins J.M. (2005), "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, **61**, 962-972.

Kim, J.K., Brick, J.M. Fuller, W.A., and Kalton, G. (2006), "On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling," Journal *of the Royal Statistical Society* B, **68**, 509-521.

Kott, P.S. (1995), "A Paradox of Multiple Imputation," *ASA Proceedings of the Survey Research Methods Section*, 380-383.

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

Rubin, D.B. (1996), "Multiple Imputation after 18+ Years," *Journal of the American Statistical Association,* **91**, 473-489.

Rubin, D.B. and Schenker, N. (1986), Multiple Imputation for Interval Estimation with Ignorable Nonresponse, *Journal of the American Statistical Association*, **81**, 366-374.