Semi-Parametric Imputation of Panel Surveys

David Judkins[1], Andrea Piesse[2], and Wen-Chau Haung[2]
[1]Abt Associates, 4550 Montgomery Avenue, Suite 800 North, Bethesda, MD 20814-3343
[2]Westat, 1600 Research Boulevard, Rockville, MD 20850-3129

Abstract
In 2007, Judkins, Krenzke, Piesse, Fan, and Haung reported on the performance of a new semi-parametric imputation algorithm designed to impute entire questionnaires with minimal human supervision while preserving important first- and second-order distributional properties. In a 2008 paper, we reported on procedures for post-imputation variance estimation to be used in conjunction with the semi-parametric imputation algorithm. In this paper, we discuss recent enhancements to handle very large longitudinal datasets for the Mental Health Treatment Study.

Key Words: Cyclic *p*-partition hotdecks

1. Introduction

Panel surveys often have very high rates of cumulative nonresponse. The number of study participants with complete records can be vanishingly small (Marker, et al., 2001). The most common solution in the past has been a mixture of hotdecks for scattered item nonresponse and weighting adjustments for attrition, with strong consideration of last observation carried forward (LOCF) for wave nonresponse (Kalton, 1986, Lepkowski, 1989, Singh, et al., 1990). Bolder thinking since then has demonstrated that larger variance reductions can be achieved, at least for targeted outcomes (Ezzati-Rice, et al., 1995) through broader use of imputation. Along these lines, recently the analysis of a large panel survey conducted as part of the Mental Health Treatment Study (MHTS) was performed using imputation for scattered item missingness, missed waves, and even attrition for those who responded at baseline and to at least two followup rounds (Frey, et al., 2011). In this paper, we report on a Monte Carlo evaluation of the imputation methodology used for the MHTS.

The methodology is a slight revision to that tested in Judkins, et al. (2007). The core of the algorithm is based on cyclic *p*-partition hotdecks (Judkins, 1997). Adapting terminology from more recent literature such as van Buuren and Groothuis-Oudshoorn (2011), the core could also be referred to as chained model-assisted hotdecks or as semi-parametric fully conditional specification. For a general discussion of hotdecks, see Andridge and Little (2010). As mentioned there, little has been proven about the asymptotic properties of hotdecks and there has been no theoretical development at all for cyclic *p*-partition hotdecks. What is known about their properties has been discovered by simulation studies and is generally encouraging. The general idea of the research presented here is to develop a complex superpopulation with challenging nonresponse patterns and then apply the methodology to samples from that superpopulation.

In the remaining sections of this paper, we document the imputation algorithm, describe the artificial population that was constructed for the evaluation, discuss performance measures, present results, and close with further discussion. One difficult issue for this type of research is the selection of a foil to place the performance of the studied method in context. We decided to use complete case analysis as the foil, by which we mean analysis based on the set of cases that have complete values for all variables required for a particular analysis. When the missing data mechanism is not MCAR (missing completely at random, as defined in Little and Rubin, 2002), it is well known that this type of analysis is easily outperformed – at least for simple statistics like marginal means. Because our simulation involves a missing data mechanism that is NMAR, it might be argued that we have picked too easy of a foil for our methodology. However, complete case analysis is still a very common approach and for complex multivariate statistics, it is not even clear that it can easily be beaten.

2. Imputation Algorithm

Let $Y_1,\ldots,Y_p$ be a collection of variables that require imputation. Let X be a vector of other variables that are never missing, such as frame variables. For each variable Y_i to be imputed, let $\wp_i = h_i\left(X, \left\{Y_{j\neq i}\right\}\right)$ be a partition of the dataset. Within each cell of $\wp_i$, cases (beggars) with missing values of Y_i are randomly matched to cases (donors) with nonmissing values of Y_i. The value from the matched donor is then imputed to the beggar. Each cell of $\wp_i$ is defined by the skip controllers[1] of Y_i and by coarsened predicted values, $\hat{Y}_i$, of Y_i. Each potential donor is used once before any is used twice. If beggars outnumber donors within a cell by a user-selected factor, then donors are sought from neighboring cells of the partition. In this search, donors with less well matched predicted values of Y_i are accepted, but donors with different skip controller values are never selected. A "sweep" involves executing this procedure once for every Y_i. Multiple sweeps are performed until either the R-squared coefficients for the models show only minimal change from sweep to sweep or until an upper limit on the number of sweeps is reached. If coarsening of $\hat{Y}_i$ is not used and if there are no skip patterns, then the algorithm is nearly equivalent to chained predictive mean matching as in Siddique and Belin (2008).

In order to reduce the need for human supervision, the predicted values of Y_i are obtained from stepwise regression models. To speed execution, the regression models are linear even for categorical variables, provided that that the categorical variables are ordered. For unordered categorical variables, a separate stepwise regression model is formed for each level of the variable, and then a k-means clustering algorithm is run on the collection of predictions to form the partition.

The use of stepwise modeling procedures allows the processing of datasets with very large values of *p*. This has been found to work well in cross-sectional surveys. However, in panel surveys, the ratio of *p* to *n* (the sample size) can be so large as to lead to serious

[1] A skip controller is a variable that determines the eligibility of a respondent for additional questions on a topic. As an example, questions about smoking habits would only be asked of smokers, so smoking status would be a skip controller for all variables about smoking habits.

problems with overfit in the stepwise modeling procedure. Such problems were noted in the preliminary imputation runs on the MHTS panel data: some variables with no theoretical relationship to Y_i nonetheless entered the model for Y_i while other potential predictors with theoretical grounds for entering the model were omitted. To reduce these problems with overfit for the MHTS, restrictions were placed on the stepwise modeling procedure. The problems were more severe for monthly time series than for quarterly time series and so stronger restrictions were placed on the stepwise algorithm for modeling of monthly series than for modeling of quarterly series. In this paper, we study only the procedure used in the MHTS for quarterly series.

The restrictions placed on the stepwise selection process are as follows:

- If there are some respondents who are eligible to answer Y_i but not Y_j, then Y_j is not allowed to enter the model for Y_i;[2] and
- If $|j-i|>1$ and Y_i and Y_j are from different time series, then Y_j is not allowed to enter the model for Y_i.

Note that these restrictions allow: all variables within a wave with consistent skip controllers to enter models for each other; all waves of a time series to enter the models for all other waves of the same series; and lagged and reverse-lagged predictions across time series. The rationale for allowing lagged predictions across series was that changes in one area of life (such as marital status) often lead to changes in other areas of life (such as emotional and financial stress). If there are L time series, w waves, no skip patterns, and no other potential predictor variables, then the addition of the second restriction reduces the number of eligible predictors from $Lw-1$ to $w+3L-4$ for bounded waves and $w+2L-3$ for the first and last waves. If the number of waves is large or a Markov assumption is reasonable, then future users might wish to consider further restricting the eligible predictors by applying the maximum lag rule within a time series as well.[3] Also note that if there is a monotone pattern of nonresponse, then there is no point in allowing Y_j to enter the model for Y_i when $j>i$, but for the MHTS quarterly series, there were strong efforts to convert apparent attritors back into respondents.

The imputation algorithm can also produce multiple imputations. This was not done for the MHTS but was done for this paper. For each multiple imputation, a completely new chain of hotdeck imputations was generated with fresh stepwise searches for variable selections. It would have been simpler instead to match several donors to each beggar with a single fixed partition, but we rejected this approach because we think that the stepwise searches could be generating a fair amount of variability in the results.

There are several tuning parameters for the imputation algorithm. The most important of these govern coarsening of predicted values when forming partitions, the number of sweeps through the p variables, and the number of multiple imputations. For this paper,

[2] Note that this restriction prevents variables from entering models for their own skip controllers. For example, smoking habits are not allowed to enter the model for smoking status. This restriction was also used in the imputation work described by Judkins, et al. (2007, 2008), and is necessary to prevent perfect models that would cause the cyclic procedure to get stuck at the initial imputed values.

[3] Something similar to this was done for the imputation of the MHTS monthly time series.

we formed 20 equal-sized portions of the sample based on $\hat{Y}_i$, conducted 10 sweeps through the variable set, and drew three multiple imputations. We think that more sweeps might have been useful, but they are computationally costly. Running the procedure on 200 draws from the superpopulation described in the next section (80 time series, arranged across 10 waves) required 120 hours of high-speed server time.

3. Monte Carlo Superpopulation

We created a superpopulation of eight related time series with skip patterns, strong cross-sectional and longitudinal correlations, and non-standard distributions. Underlying the eight series was a latent nonstochastic process for each person that was driven by a random starting vector and random transition matrix. The process is perhaps best envisioned as a propensity vector for a five-level categorical variable. For person *i* at wave *j*, the latent propensity vector was generated as

$$\mathbf{L}_{ij} = \begin{bmatrix} L_{ij1} \\ L_{ij2} \\ L_{ij3} \\ L_{ij4} \\ L_{ij5} \end{bmatrix} = \begin{bmatrix} L_{i,j-1,1} \\ L_{i,j-1,2} \\ L_{i,j-1,3} \\ L_{i,j-1,4} \\ L_{i,j-1,5} \end{bmatrix}^T \begin{bmatrix} T_{i1}^T \\ T_{i2}^T \\ T_{i3}^T \\ T_{i4}^T \\ T_{i5}^T \end{bmatrix} \quad \text{for } j > 1,$$

where

$$\begin{bmatrix} L_{i11} \\ L_{i12} \\ L_{i13} \\ L_{i14} \\ L_{i15} \end{bmatrix} = Dirichlet\left(0.05 \begin{bmatrix} .05 \\ .10 \\ .70 \\ .10 \\ .05 \end{bmatrix} \right)$$

and

$$T_{i1} = Dirichlet\left(0.5 \begin{bmatrix} 0.900 \\ 0.070 \\ 0.020 \\ 0.009 \\ 0.001 \end{bmatrix} \right), \quad T_{i2} = Dirichlet\left(0.5 \begin{bmatrix} 0.30 \\ 0.50 \\ 0.15 \\ 0.04 \\ 0.01 \end{bmatrix} \right), \quad T_{i3} = Dirichlet\left(0.5 \begin{bmatrix} 0.01 \\ 0.20 \\ 0.40 \\ 0.35 \\ 0.04 \end{bmatrix} \right),$$

$$T_{i4} = Dirichlet\left(0.5 \begin{bmatrix} 0.001 \\ 0.009 \\ 0.100 \\ 0.700 \\ 0.190 \end{bmatrix} \right), \quad T_{i5} = Dirichlet\left(0.5 \begin{bmatrix} 0.001 \\ 0.001 \\ 0.048 \\ 0.150 \\ 0.800 \end{bmatrix} \right).$$

The eight manifest time series were then generated as follows. An ordinal variable for wellbeing was generated as

$$Wellbeing_{ij} \sim M\left(1,\ L_{ij1},\ldots,L_{ij5}\right).$$

A binary substance abuse indicator was generated as

$$Abuse_{ij} \sim B\left(1,\ \frac{1}{1+\exp\left(-\begin{bmatrix}30 & 3 & 0 & -3 & -20\end{bmatrix}\mathbf{L}_{ij}\right)}\right).$$

A binary prison indicator was generated as an absorbing event with the following hazard function:

$$h_{ij} = \begin{cases} 0 & \text{for } j = 1, \\ \begin{bmatrix}0.3 & 0.1 & 0.01 & 0.001 & 0.001\end{bmatrix}\mathbf{L}_{ij} & \text{for } j > 1. \end{cases}$$

Mental health was generated as

$$MHealth_{ij} \sim N\left(\begin{bmatrix}-2 & -1 & 0 & 1 & 2\end{bmatrix}\mathbf{L}_{ij},\ 1\right) - Prison_{ij}.$$

Labor force status (1=employed, 2=unemployed, 3=not in labor force) was generated as

$$\begin{aligned} Y_{ij} &= 3 && \text{if } Prison_{ij} = 1, \\ Y_{ij} &\sim M\left(1,\ \begin{bmatrix}0.01 & 0.20 & 0.50 & 0.70 & 0.70 \\ 0.25 & 0.20 & 0.15 & 0.10 & 0.02 \\ 0.74 & 0.60 & 0.45 & 0.20 & 0.28\end{bmatrix}\mathbf{L}_{ij}\right) && \text{otherwise.} \end{aligned}$$

Number of cigarettes per day was generated as a two-step process, with one random variable to determine smoking status (smoker or not) and a second to conditionally determine number of cigarettes per day:

$$C_{ij} = Any_{ij} Num_{ij}.$$

At wave 1, smoking status was generated as

$$Any_{i1} \sim B\left(1,\ \begin{bmatrix}0.7 & 0.5 & 0.3 & 0.15 & 0.05\end{bmatrix}\mathbf{L}_{i1}\right).$$

At subsequent waves, persistent smoking was generated as

$$Any_{ij} \left| \left(Any_{i,j-1} = 1\right) \sim B\left(1,\ \max\left(0.01,\ \min\left(0.99,\ \begin{bmatrix}0.99 & 0.98 & 0.97 & 0.96 & 0.95\end{bmatrix}\mathbf{L}_{i,j-1}\right)\right)\right)\right.$$

and uptake was generated as

$$Any_{ij} \left| \left(Any_{i,j-1} = 0\right) \sim B\left(1,\ \max\left(0.01,\ \min\left(0.99,\ \begin{bmatrix}0.1 & 0.05 & 0.01 & 0.01 & 0.005\end{bmatrix}\mathbf{L}_{i,j-1}\right)\right)\right).\right.$$

For the first wave, the daily cigarette consumption for smokers was generated as

$$\log\left(Num_{i1}\middle|Any_{i1}=1\right)\sim N\left(\left[1.6\quad 1.8\quad 2\quad 2.5\quad 3\right]\mathbf{L}_{i1},\ 0.25\right).$$

For subsequent waves, the daily cigarette consumption for smokers was generated as

$$\log\left(Num_{ij}\middle|Any_{ij}=1\right)\sim N\left(\left[1.6\quad 1.8\quad 2\quad 2.5\quad 3\right]\mathbf{L}_{i1},\ 0.25\right)\quad \text{if } Any_{i,j-1}=0,$$

$$Num_{ij}\middle|\left(Any_{ij}=1\right)\sim \Gamma\left(\left(\frac{Num_{i,j-1}}{3}\right)^2,\min\left\{1,\frac{9}{Num_{i,j-1}}\right\}\right)\qquad \text{if } Any_{i,j-1}=1.$$

Hours worked per week for the employed was constructed with an usual distribution, generated as

$$Hours_{ij}=\text{round}\left\{1+\frac{39}{1+\exp\left(\left[10\quad 7\quad 3\quad 0\quad -2\right]\mathbf{L}_{ij}-8U_{ij}\right)}\right\},$$

where U_{ij} was randomly drawn from the uniform distribution between 0 and 1. Hours worked per week for the unemployed and those not in the labor force were, of course, set to zero.

Income was generated as a contaminated heteroscedastic normal variable. Income outliers were generated with a probability of 0.005 and their income was multiplied by factor of 1.5. Otherwise, income was generated to be heteroscedastic normal as

$$Income_{ij}\sim N\left(\left[1{,}000\quad 5{,}000\quad 20{,}000\quad 80{,}000\quad 200{,}000\right]\mathbf{L}_{ij},\right.$$
$$\left.\left\{\left[200\quad 1{,}000\quad 4{,}000\quad 16{,}000\quad 40{,}000\right]\mathbf{L}_{ij}\right\}^2\right).$$

If the resulting income amount was negative, it was rounded up to $0. Also, for the unemployed, the resulting income amount was halved, and for those in prison, the amount was reduced by factor of 10.

Item nonresponse was generated as completely at random with a rate of 5% for all variables. Wave nonresponse was also generated as completely at random with a rate of 4%. Attrition was generated as ignorable given wellbeing and prison at last wave, but these were, of course, not always observed. So attrition is nonignorable, but of a form that should be handled reasonably well by procedures that assume ignorable nonresponse. Attrition was 3% per wave but the logit of the probability of attrition was

$$\text{logit}\left[\Pr\left\{Attrit_{ij}\right\}\right]=-3.48+3\left(1-\frac{Wellbeing_{i,j-1}}{3}\right)+1.2\left(Prison_{i.j-1}-0.7\right).$$

The wave 1 sample size was set to 2,000 and the number of waves was set to 10.

4. Performance Measures

We envision the imputation process as being carried out by a data publisher rather than by an expert analyst. The goal of this approach to imputation is to create a rectangular dataset that can meet the needs of future "journeyman" or even novice analysts. An expert analyst working on a small subset of variables might prefer to try the Bayesian methods in Chapter 21 of Gelman, et al. (2004) or some of the frequentist methods in Molenberghs and Kenward (2007), but these methods are beyond the skill levels of most analysts. Fay (1993) demonstrated how difficult it can be to support unplanned analyses well. Given the way that partitions are built and beggars are matched to donors by our imputation method, one can expect that examinations of univariate statistics will be well supported and hope that the same will be true for bivariate statistics; however, it would obviously be too much to hope that three-way or higher relationships will be well supported, other than those induced by skip patterns.

Accordingly, we evaluated the performance of the imputation algorithm for marginal means, standard deviations, skewnesses, and kurtoses and for conditional means of binary and ordinal variables given (other) categorical variables. Two common analyses of panel datasets are growth curve modeling and hazard modeling, so we also evaluated the performance of the imputation for supporting these analyses. For continuous variables (wellbeing, mental health, daily cigarette consumption, hours worked per week, and income), we fit models of the form

$$
\begin{aligned}
&Y_{ij} = \mu + \beta j + u_i + b_i j + e_{ij} \\
&u_i \sim N\left(0,\ \tau^2\right) \\
&\mathrm{Cov}\left(u_i, u_j\right) = 0 \quad \text{for } i \neq j \\
&b_i \sim N\left(0,\ \varphi^2\right) \\
&\mathrm{Cov}\left(b_i, b_j\right) = 0 \quad \text{for } i \neq j \\
&\mathrm{Cov}\left(u_i, b_j\right) = 0 \quad \forall i, j \\
&e_{ij} \sim N\left(0,\ \sigma^2\right) \\
&\mathrm{Cov}\left(e_{ij}, e_{i'k}\right) = \begin{cases} 0 & \text{for } i \neq i', \\ \rho^{|k-j|}\sigma^2 & \text{for } i = i'. \end{cases}
\end{aligned}
$$

This is not how any of the variables were generated, so the models are wrong but possibly useful (as are all models, as famously noted by G.E.P. Box). We evaluated our imputation method by comparing the parameter estimates from fitting the model on the complete data to the parameter estimates from fitting the model on the imputed data. We focused on the estimation of β / σ, φ / σ, ρ, and $\delta = \dfrac{\tau^2}{\tau^2 + \sigma^2}$.

For the binary variables except prison (abuse, smoking status, employed, unemployed, and not in the labor force), we fit models of the form:

$$
\begin{aligned}
&Y_{ij} \sim B\left(1,\ \pi_{ij}\right) \\
&\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \mu + \beta j + u_i + b_i j \\
&u_i \sim N\left(0,\ \tau^2\right) \\
&\mathrm{Cov}\left(u_i, u_j\right) = 0 \quad \text{for } i \neq j \\
&b_i \sim N\left(0,\ \varphi^2\right) \\
&\mathrm{Cov}\left(b_i, b_j\right) = 0 \quad \text{for } i \neq j \\
&\mathrm{Cov}\left(u_i, b_j\right) = 0 \quad \forall i, j.
\end{aligned}
$$

Again, these models are "wrong" but potentially useful. As with the linear mixed models, we evaluated the imputation algorithm by comparing the results of fitting the model on the complete data and on the imputed data. For the binary growth models, we focused on the estimation of $\beta/1.65$, $\varphi/1.65$, and $\delta = \frac{\tau^2}{\tau^2 + 3.29}$. (The 1.65 comes from a suggestion by Sir David Cox favorably evaluated for meta-analysis by Sánchez-Meca, et al., 2003, and the $3.29 = \pi^2/3$ comes from a suggestion by Donald Hedeker built into his MIXOR program, Hedeker and Gibbons, 1996.)

For prison (created as an absorbing binary variable), we fit a Cox proportional hazards model with a single covariate (abuse at wave 1) and no random effects.

We computed nominal 95-percent confidence intervals for the marginal means (of 11 variables – the seven ordered variables, binary indicators for the three levels of the unordered categorical variable, and the binary indicator for smoking status), conditional means (of the other ten variables given each categorical variable), growth rates (the β coefficients from 11 growth models), and the hazard rate of prison given abuse at wave 1. These nominal confidence intervals were calculated using Donald Rubin's standard formula for post-imputation variance (equation 2.2 in Rubin, 1996) and infinite degrees of freedom. We then computed empirical coverage rates for these nominal confidence intervals where success was including the cross-replicate average of the same statistic based on analysis of complete datasets.

To describe this using formulae, let *r* index repeated draws from the superpopulation and *t* index multiple imputations. Let $\hat{\theta}_{Fr}$ represent the estimated parameter on the full sample with complete response, $\hat{\theta}_{Irt}$ represent the estimated parameter on the imputed dataset, $\hat{\theta}_{Cr}$ represent the estimated parameter based on cases that have complete values for all variables required to compute the statistic, and $\hat{Q}_{Fr}$, $\hat{Q}_{Irt}$, and $\hat{Q}_{Cr}$, respectively, represent the naïve variance estimates for these point estimates. Limits for nominal confidence intervals based on the imputed data were calculated as $\frac{1}{T}\sum_t \hat{\theta}_{Irt} \pm 1.96\sqrt{\frac{1}{T}\sum_t \hat{Q}_{Irt} + \frac{T+1}{T}\frac{1}{T-1}\sum_t\left(\hat{\theta}_{Irt} - \frac{1}{T}\sum_s \hat{\theta}_{Irs}\right)^2}$. Limits for nominal confidence intervals based on the complete case analysis were calculated as

$\hat{\theta}_{Cr} \pm 1.96\sqrt{\hat{Q}_{Cr}}$. The confidence intervals were classified as successful if they contained $\frac{1}{R}\sum_{r}\hat{\theta}_{Fr}$.

Bias and root mean square error (RMSE) were also computed for all statistics of interest. These were calculated as

$$B_I = \frac{1}{RT}\sum_{r,t}\hat{\theta}_{Irt} - \frac{1}{R}\sum_{r}\hat{\theta}_{Fr}\ ,\ \ RMSE_I = \sqrt{\frac{1}{R-1}\sum_{r}\left(\frac{1}{T}\sum_{t}\hat{\theta}_{Irt} - \frac{1}{R}\sum_{s}\hat{\theta}_{Fs}\right)^2}\ ,$$

$$B_C = \frac{1}{R}\sum_{r}\hat{\theta}_{Cr} - \frac{1}{R}\sum_{r}\hat{\theta}_{Fr}\ ,\ \text{and}\ \ RMSE_C = \sqrt{\frac{1}{R-1}\sum_{r}\left(\hat{\theta}_{Cr} - \frac{1}{R}\sum_{s}\hat{\theta}_{Fs}\right)^2}$$

Note that it might also have been interesting to calculate an alternative measure of root mean square error where the pivot was the full sample statistic from the draw from the superpopulation. However, we did not compute that measure. We felt that accuracy in estimating the fundamental quantity of interest, $\sum_{s}\hat{\theta}_{Fs}$, was more important than approximating the full sample statistic, $\hat{\theta}_{Fr}$, on every draw.

5. Results

Due to the computational intensity of the imputation algorithm, performance was assessed on only 200 draws from the superpopulation. Figure 1 shows empirical coverage rates for nominal 95-percent confidence intervals, for four types of analysis by two approaches to handling missing data. For marginal means, our imputation procedure dramatically outperforms complete case analysis, but in absolute terms, performance still leaves something to be desired. Empirical coverage rates for the 11 marginal means varied from 55 to 94 percent. These coverage problems were caused by a combination of bias and underestimation of variance. Our imputation procedure also outperformed complete case analysis for conditional means in two-way analyses, however the advantage was not as strong presumably because the missing data mechanism did not create as much bias in conditional means. For estimation of growth rates, our procedure again dramatically outperformed complete case analysis, but did not perform as well as desired in absolute terms. The two methods both did well for hazard analysis.

Where the complete case methodology had poor coverage, it was because of bias. Figures 2 and 3 show various biases. For marginal means, standard deviations, and skew and kurtosis measures, our method is virtually unbiased despite the NMAR attrition, while complete case analyses are badly biased. The results for wave 10 (W10) correlations are particularly satisfying in that imputation methods are well known for causing attenuation of correlation, but our method actually led to less attenuation on this superpopulation than did complete case analysis. Our method also performed very well for average growth rates (upper left panel of Figure 3 shows β/σ) and for the standard deviation of personal growth rates (lower right panel of Figure 3 shows φ/σ). For intraclass correlations (ICCs) and autoregressive correlations, the imputation method was generally okay, but for one of the outcomes, it borrowed too much strength from distant waves

rather than neighboring waves and thereby ended up overestimating the ICC and underestimating the autoregressive correlation.

Figures 4 and 5 show root mean square errors. The pattern of results is consistent with those for biases. Our method has much lower RMSE than complete case analysis for marginal statistics, cross-sectional correlations, average growth rates, and the standard deviation of personal growth rates. The two methods were roughly tied for conditional means, and complete case analysis was better for estimating both intraclass correlations and autoregressive correlations.

6. Discussion

The superpopulation crafted for this research would, we believe, pose a serious challenge for any imputation system. We hope that others will take up the challenge and test their systems against it. The SAS code for generating and analyzing the population is available from the authors. The code for our imputation system is a proprietary product of Westat and is not available for sharing. However, in this and previous papers, we have shared the core algorithmic details.

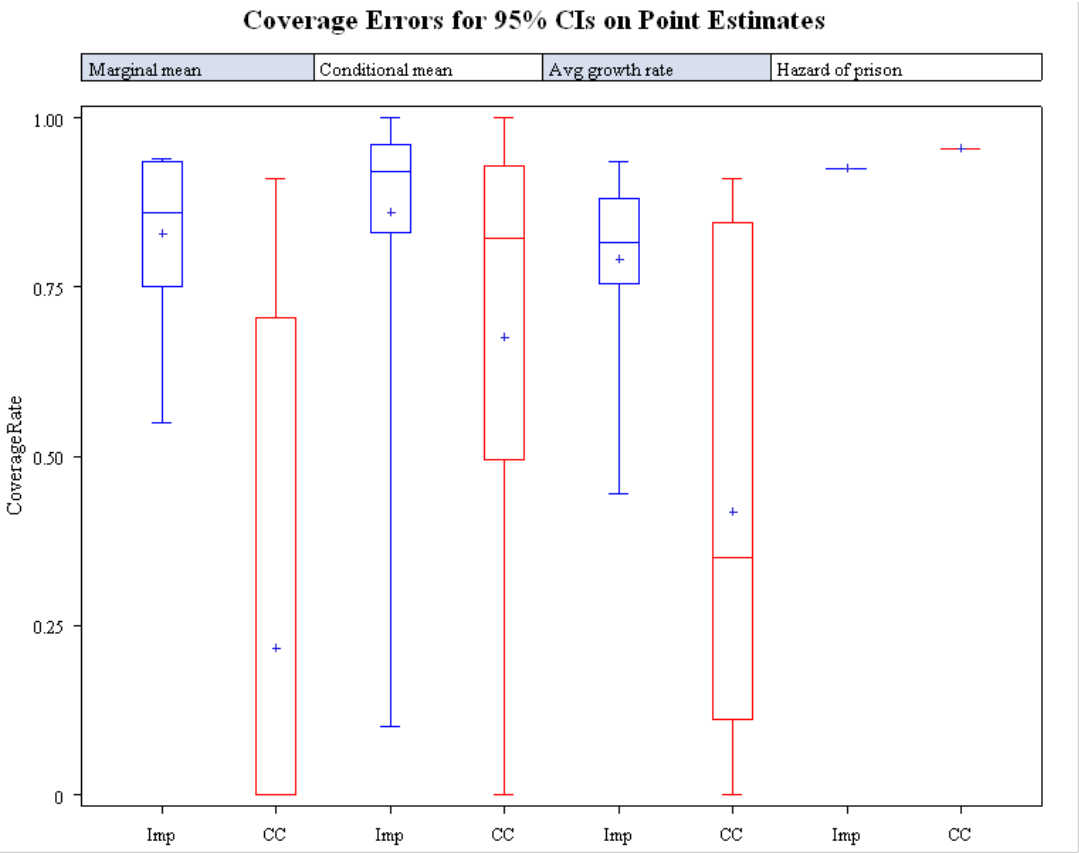

Figure 1: Coverage rates by type of statistic and approach to missing data

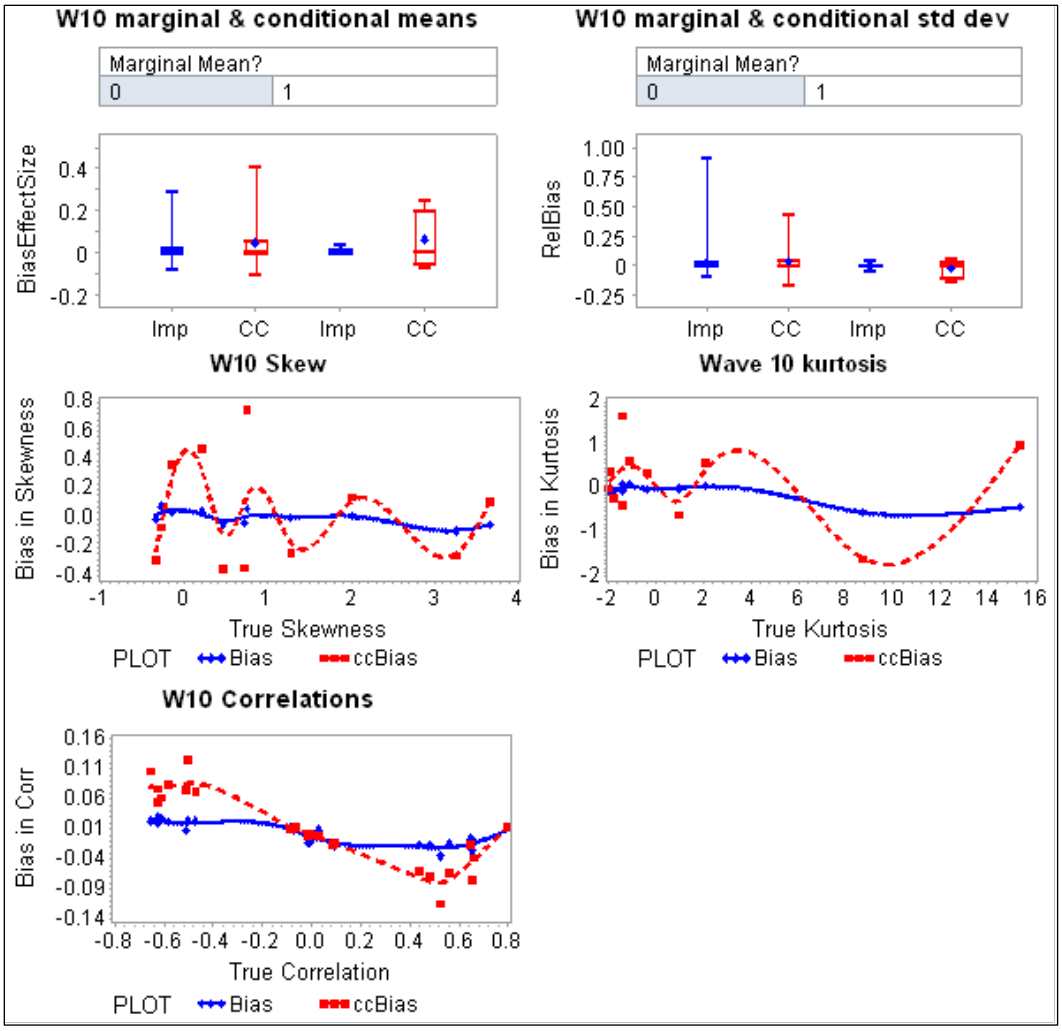

Figure 2: Biases in cross-sectional parameters by type of statistic and approach to missing data

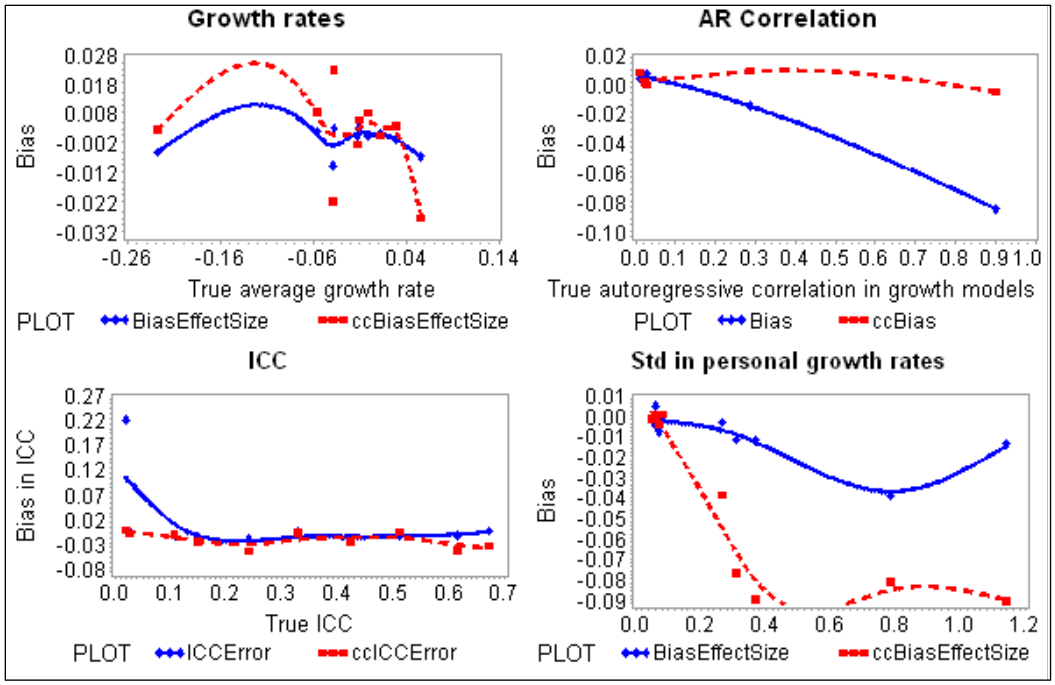

Figure 3: Biases in longitudinal parameters by type of statistic and approach to missing data

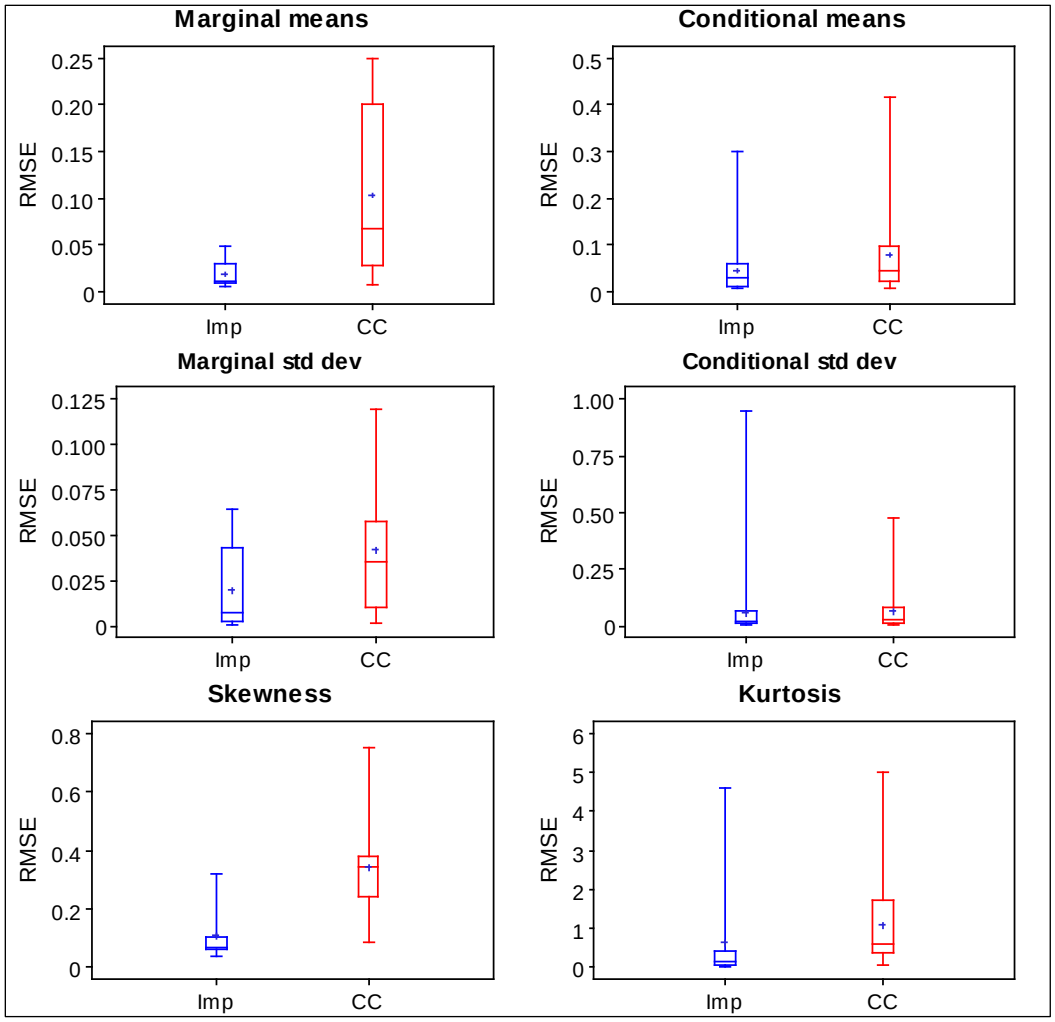

Figure 4: Root mean square errors of cross-sectional statistics at wave 10 by type of statistic and approach to missing data

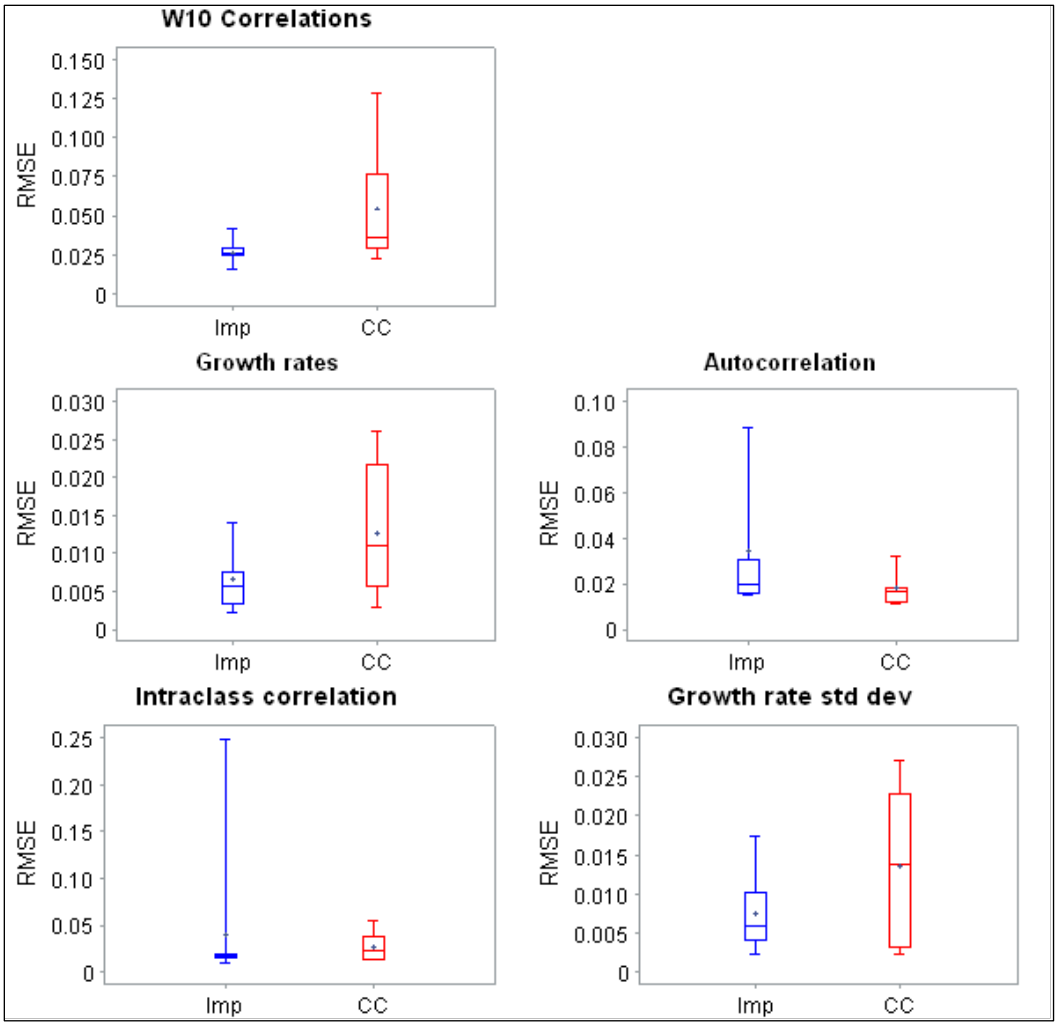

Figure 5: Root mean square errors of cross-sectional correlations at wave 10 and of longitudinal statistics by type of statistic and approach to missing data

References

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

Ezzati-Rice, T., Johnson, W., Khare, M., Little, R.J.A., Rubin, D., and Schafer, J. (1995). A simulation study to evaluate the performance of model-based multiple imputations in NCHS health examination surveys. *Proc. 1995 Annual Research Conf., U.S. Bureau of the Census*, 257-266.

Fay, R.E. (1993). Valid inferences from imputed survey data. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 227-232.

Frey, W.D., Drake, R.E., Bond, G.R., Miller, A.L., Goldman, H.H., Salkever, D.S., and Holsenbeck, S. (2011). *Mental Health Treatment Study: Final Report*. Rockville, MD: Westat.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis, 2nd ed.* Boca Raton: Chapman and Hall.

Hedeker, D. and Gibbons, R.D. (1996). MIXOR: a computer program for mixed-effects ordinal probit and logistic regression analysis. *Computer Methods and Programs in Biomedicine*, **49**, 157-176.

Judkins, D. R. (1997). Imputing for Swiss cheese patterns of missing data. *Proceedings of Statistics Canada Symposium 97, New Directions in Surveys and Censuses*, 143-148.

Judkins, D., Krenzke, T., Piesse, A., Fan, Z., and Haung, W.C. (2007). Preservation of skip patterns and covariance structure through semi-parametric whole-questionnaire imputation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 3211-3218.

Judkins, D., Piesse, A., and Krenzke, T. (2008). Multiple semi-parametric imputation. *Proceedings of the Joint Statistical Meetings* [CD-ROM], pp. 48-58. Alexandria, VA: American Statistical Association.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, **2**, 303-314.

Lepkowski, J.M. (1989). Treatment of wave nonresponse in panel surveys, in *Panel Surveys* (Eds. D. Kasprzyk, G. Kalton, and M.P. Singh), 348-374.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data, 2nd ed*. Hoboken, NJ: John Wiley and Sons.

Marker, D.A., Judkins, D.R., and Winglee, M. (2001). Large-scale imputation for complex surveys, in *Survey Nonresponse* (Eds. R.M. Groves, D.A. Dillman, E.L. Eltinge, and R.J.A. Little). New York: Wiley.

Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies*. Chichester: Wiley.

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

Sánchez-Meca, J., Marín-Martínez, F., and Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, **8**, 448-467.

Siddique J. and Belin T.R. (2008). Multiple imputation using an iterative hot-deck with distance-based donor selection. *Statistics in Medicine*, **27**, 83–102.

Singh, R., Huggins, V. and Kasprzyk, D. (1990). *Handling Single Wave Nonresponse in Panel Surveys.* SIPP Working Paper #114. Washington DC: U.S. Census Bureau.