Stat Meth Appl (2008) 17:351–372
DOI 10.1007/s10260-007-0090-4

ORIGINAL ARTICLE

A cautionary case study of approaches to the treatment of missing data

Christopher Paul · William M. Mason · Daniel McCaffrey · Sarah A. Fox

Accepted: 11 December 2007 / Published online: 8 January 2008

Abstract This article presents findings from a case study of different approaches to the treatment of missing data. Simulations based on data from the Los Angeles Mammography Promotion in Churches Program (LAMP) led the authors to the following cautionary conclusions about the treatment of missing data: (1) Automated selection of the imputation model in the use of full Bayesian multiple imputation can lead to unexpected bias in coefficients of substantive models. (2) Under conditions that occur in actual data, casewise deletion can perform less well than we were led to expect by the existing literature. (3) Relatively unsophisticated imputations, such as

The research reported here was partially supported by National Institutes of Health, National Cancer Institute, R01 CA65879 (SAF). We thank Nicholas Wolfinger, Naihua Duan, John Adams, John Fox, and the anonymous referees for their thoughtful comments on earlier drafts. The responsibility for any remaining errors is ours alone. Benjamin Stein was exceptionally helpful in orchestrating the simulations at the labs of UCLA Social Science Computing. Michael Mitchell of the UCLA Academic Technology Services Statistical Consulting Group artfully created Fig. 1 using the Stata graphics language; we are most grateful.

C. Paul (✉) · D. McCaffrey
RAND, 4570 Fifth Ave., Suite 600, Pittsburgh, PA 15213, USA
e-mail: cpaul@rand.org

D. McCaffrey
e-mail: Daniel_McCaffrey@rand.org

W. M. Mason
California Center for Population Research, University of California, Los Angeles,
4284 Public Policy Building, PO Box 951484, Los Angeles, CA 90095, USA
e-mail: masonwm@ucla.edu

S. A. Fox
Department of Medicine, Division of General Internal Medicine and Health Services Research,
University of California, Los Angeles, 1100 Glendon Ave., Suite 2010, Los Angeles, CA 90024, USA
e-mail: sfox@mednet.ucla.edu

mean imputation and conditional mean imputation, performed better than the technical literature led us to expect. (4) To underscore points (1), (2), and (3), the article concludes that imputation models *are* substantive models, and require the same caution with respect to specificity and calculability.

Keywords Missing data · Imputation · Multiple imputation · Casewise deletion

1 Introduction

This article is a case study of different approaches to the treatment of missing data. Specifically, the following techniques: casewise deletion, weighted casewise deletion, mean imputation, mean imputation with a dummy for missingness, conditional mean imputation, hotdeck imputation, approximate Bayesian bootstrap (ABB) multiple imputation, and full Bayesian multiple imputation. This article reaches the following cautionary conclusions about the treatment of missing data: (1) Compared to other missing data techniques we tried, automated selection of the imputation model in the use of full Bayesian multiple imputation can lead to greater bias in coefficients of substantive models. (2) Under conditions that occur in actual data, casewise deletion can perform less well than we were led to expect by the existing literature (specifically, Allison 2001). We find that relatively minor violations of assumptions can produce this result. (3) Relatively unsophisticated imputations, such as mean imputation and conditional mean imputation, performed better than the technical literature led us to expect. (4) To underscore points (1), (2), and (3), we conclude that imputation models *are* substantive models, and require the same caution with respect to specificity and calculability.

Why consider and evaluate multiple alternative techniques for dealing with missing data, given that Allison (2001, pp. 5–6) indicates that none of the widely used "conventional" methods for dealing with missing data is any better than listwise (casewise) deletion?[1] Because many of these other techniques are quite common in practice. While casewise deletion is the default treatment of missing data by most software and many analysts, mean imputation with a dummy for missingness is also a common practice. The US Bureau of the Census still uses hotdeck imputation. And, despite the apparent theoretical superiority of multiple imputation, practice remains heterogeneous.

This case study has results that are contrary to the "received wisdom" on the treatment of missing data. Because these "bad" things happened once to us in our case study of real data, it is not impossible that they will happen to you in your analysis of other real data. We hope that this article is successful at illustrating some of the pitfalls in the application of missingness techniques that await even the wary. The target audience is the journeyman data analyst; many of the issues raised and discussed here may be clear and obvious to a veteran statistician.

[1] "Conventional" methods are essentially those that predate the multiple imputation and maximum likelihood approaches to the treatment of missing data.

2 Data and core analysis

The data used for this case study and the associated simulation analysis come from a real missing data problem: Twenty-eight percent of responses to a household income question were missing in a survey for the Los Angeles Mammography Promotion in Churches Program (LAMP) to whose design we contributed (Fox et al. 1998). Since economic well-being was thought to be important for the topic that was the focus of the survey—compliance with guidelines for regular mammography screening among women in the United States—there were grounds for concern with the quantity of missing responses to the household income question. Fox et al. (1998) estimated screening guideline compliance as a function of household income and other covariates using the "approximate Bayesian bootstrap" (Rubin and Schenker 1986, 1991) to compensate for missingness on household income.

In the original analysis, all variables are discrete and most, including the response, are dichotomous. Estimation is carried out with logistic regression. A respondent is considered "compliant" if she had a mammogram within the 24 months prior to the baseline interview and another within the 24 months prior to that most recent mammogram, and is considered "noncompliant" otherwise. Our list of regressors[2] consists of dummy variables (coded one in the presence of the stated condition and zero otherwise) for whether the respondent is (1) Hispanic; (2) has medical insurance of any kind; (3) is married or living with a partner; (4) has been seeing the same doctor for a year or more; (5) is a high school graduate; (6) lives in a household with annual income greater than $10,000 per year; (7) has a doctor she regards as enthusiastic about mammography; and a trichotomous dummy variable classification for (8) whether the respondent's doctor is Asian, Hispanic, or belongs to another race/ethnicity group.

Deletion of a respondent if information is missing on any variable in the model, including the response variable (casewise deletion), reduces the sample size to 857 cases, or 56 percent of the total sample. This is the result of a great deal of missingness on a single covariate, and the cumulation of a low degree of missingness on the response and remaining covariates. As noted earlier, 28 percent of respondents refused to disclose their household annual income—by far the highest level of missingness in the data set.[3] The next highest level of missingness (seven percent) occurs for the response variable, mammography screening compliance. A number of respondents could not recall their mammography history in detail sufficient to allow discernment of their compliance status.

Discarding respondents who are missing on mammography compliance or any covariate in the logistic regression model *except* household income results in a data set of 1,119 individuals, or 76 percent of the total sample. For present purposes we define this subsample of 1,119 individuals to be the working sample of interest. In the working sample, 23 percent (262 respondents out of 1,119) refused or were unable to answer the household income question.

[2] See Fox et al. (1998) for details and Breen and Kessler (1994) and Fox et al. (1994) for additional justification.

[3] Respondents were given 10 household income intervals with a top code of "$25,000 or more" from which to select. In the computations presented here, we treat "don't know" and "refused" as missing.

Missingness on household income provides the point of departure into our exploration of techniques for dealing with missingness. Our initial calculations on the actual LAMP data demonstrate the effects on the logistic regression for mammography compliance of various treatments of missing household income. The closely related simulated data enable examination of the performance of different missingness techniques across various assumptions about the *nature* of the missingness process.

3 Missingness techniques and mechanisms

Techniques for dealing with missingness can be evaluated for the extent to which they induce coefficient (b) and standard error [$SE(b)$] bias, where "bias" is specified relative to samples with no missing data, and for the extent to which they increase or decrease the sampling variance of the coefficients [$Var(b)$]. The performance of a missingness technique as defined by these three characteristics depends on the mechanism of missingness present in a given body of data. Note that the use of the "bias" concept assumes that the substantive model is *perfectly* specified. For the case considered in this article—missingness on a single regressor—when we assert that a substantive regression is perfectly specified, we mean that it has the correct error distribution and functional form; that it excludes no relevant regressors (whether in the data or not); that it includes all necessary interactions between regressors; and that it contains no regressor with measurement error. In actual research practice, data analysts are unlikely to know whether a substantive model is perfectly specified, and it strains credulity to suggest that most are. Although we believe the model used for the example in this article is plausible, we do not know if it is perfectly specified, and our simulation analyses reveal that probably it is not.

3.1 Mechanisms of missingness

This section reviews the typology of mechanisms of missingness for non-specialists.

Let Y denote the response variable for mammography compliance. Let X denote the dichotomy for household income, and let Z denote not only the covariates in the logistic regression model, but all variables in the data other than Y and X. Mechanisms of missingness can be defined with reference to a missingness model—a model for the probability that a respondent is missing on X. Let $R_i = 1$ if the ith respondent is missing on X, and let $R_i = 0$ if the ith respondent provides a valid response on X. Three mechanisms of missingness are:

1. The probability that $R_i = 1$ is independent of Y, Z, and X itself;
2. The probability that $R_i = 1$, conditional on Y and Z, is independent of X;
3. The probability that $R_i = 1$, conditional on Y and Z, depends on X.

The first missingness mechanism is known as *m*issing *c*ompletely *a*t *r*andom (MCAR). If household income is MCAR, then the observed values are a random sample of all values (observed and unobserved).

The second missingness mechanism is known as *m*issing *a*t *r*andom (MAR). Missingness on household income is MAR if it does not depend on the actual value

(even if unobserved) of household income itself once other variables in the data (some subset of Y and Z) are controlled.

Missing completely at random is a special case of missing at random. With MAR, missingness has a purely random component and a systematic component that depends on some variable(s) in the data set, but not on the actual values of the variable with missingness. With MCAR, the missingness has only a purely random component.

When the MCAR and MAR assumptions (missingness has a random component; missingness does not depend on X, conditional on Y and Z) are combined with the technical assumption of "parameter distinctness" (Schafer 1997a, p. 11; Little and Rubin 2002; Rubin 1987), the missingness mechanism is termed "ignorable." The ignorability assumption is a necessary condition for modeling substantive relationships in the data set separately from modeling missingness per se, or imputing missing values.[4]

The third missingness mechanism is known as *m*issing *n*ot *a*t *r*andom (MNAR), also referred to as "nonignorable" in much published research. If missingness on household income is MNAR, it depends on the actual level of household income and potentially other variables as well. Note that MNAR does *not* mean that missingness lacks a random component, only that its systematic component is a function of the actual values of the variable with missingness.[5]

In actual practice, it is difficult to know whether missingness is ignorable, especially with cross-sectional data, and it seems a plausible conjecture that some degree of nonignorability in missingness processes is common.[6] Here, as in many other situations, a continuum is probably more realistic than an "all or none" typology, and, presumably, a little nonignorability differs from a lot. The assumption of nonignorability in the missingness model parallels the assumption that in the substantive model the covariates and disturbance are orthogonal. Most researchers (implicitly) argue that if the orthogonality assumption is not perfectly satisfied by their substantive model, then the distortion caused by nonorthogonality is not so great as to obscure the pattern of interest. For this reason, in the simulations introduced in later sections we allow for differing degrees of nonignorability.

3.2 Missingness techniques

This section briefly discusses the eight missing data techniques considered in the article: casewise deletion, weighted casewise deletion, mean imputation, mean imputation with a dummy for missingness, conditional mean imputation, hotdeck, approximate Bayesian bootstrap multiple imputation, and full Bayesian multiple imputation. These

[4] For other conditions, see Schafer (1997a, p. 10).

[5] When MNAR is considered by the analyst to be the overriding feature of missingness for a specific variable, the difficulty is generally viewed as a sample selection problem, in which case the missingness model and the substantive model are inseparable (e.g., Heckman 1976, 1979). The complexities engendered by solutions to missingness under nonignorability are beyond the scope of this article.

[6] Groves et al. (2000) document an instance of nonignorability using a two-wave panel study; Carpenter et al. (2007) believe MNAR missingness to be sufficiently common to advocate sensitivity analysis to diagnose its possible impact when using multiple imputation.

techniques are the techniques that were considered for use in the original empirical analysis on which this case study draws and can thus be considered part of the "realness" of the case study. While certainly not exhaustively representative of currently used approaches to the treatment of missing data (notably absent is the maximum likelihood approach advocated by Allison 2001), these eight techniques broadly represent existing practice.

3.2.1 Casewise deletion

The standard treatment of missing data in most statistical packages—and hence the default treatment for most analysts—is the deletion of any case containing missing data on one or more of the variables used in the analysis. Use of this approach assumes that either (a) the missingness and imputation models have no covariates (missingness is MCAR) or (b) that the substantive model is perfectly specified, *and* that the missingness mechanism is a special case of MAR or MNAR in which Y is not a covariate in the missingness model (equivalently, Y is uncorrelated with missingness on X).[7] If either assumption is satisfied, then unbiased coefficient estimates may be obtained without imputation. Also, the coefficient standard errors will be valid for a sample of reduced size.

Based on these properties, Allison (2001) argues for the superiority of casewise deletion over other "conventional" approaches (by which he means all approaches that are not maximum likelihood or multiple imputation approaches). He observes that casewise deletion gives unbiased coefficient estimates when missingness is MCAR, and under certain very specific situations of MAR and even MNAR as well. He further observes that conventional imputation methods are not guaranteed to provide unbiased coefficient estimates in the situations that casewise deletion is, but that they *are* guaranteed to produce underestimated standard errors because they fail to adjust for the uncertainty associated with the missing values that have been imputed (this is what multiple imputation corrects for—see Sect. 3.2.7, below).

While the mathematics that he bases these claims on are correct, we take issue with the implied frequency of occurrence of the conditions under which casewise deletion is clearly superior to other conventional techniques. As the simulation analyses in Sect. 5 show, this case study happens to be one case in which casewise deletion is *not* superior. Ambler and Omar (2007) provide another example where casewise deletion produces unreliable predictions in simulations with real data. We have found no systematic evidence in the literature on missing data techniques regarding the frequency with which assumptions needed for case deletion have been checked and found valid, in data analyzed in the social sciences.

[7] The discussion of the OLS-specific Theorem 2.1 in Jones (1996) provides the basis for this assertion. Allison (2001, p. 7, footnote 1) offers a proof that is valid for any regression procedure, including logistic regression. Citing Vach (1994), Allison (2001, p. 7) also notes that if missingness is entirely determined by the dependent variable of the substantive model (Y), logistic regression with listwise deletion will result in consistent covariate coefficient estimates. In this instance, missingness conforms to the case-control design, for which the consistency result is well established (e.g., Farewell 1979). The applicability of the result in a specific instance hinges on the validity of the assumed logistic functional form (Xie and Manski 1989).

3.2.2 Weighted casewise deletion

Weighted casewise deletion extends the range of MAR models under which unbiased coefficient estimation in the substantive model can be achieved. Specifically, if the substantive model is perfectly specified, and if missing data are MAR, and if missingness is correlated with Y, then weighted casewise deletion can result in unbiased coefficient estimation of the substantive model (Brick and Kalton 1996). Nonresponse weighting increases the weight of complete cases to represent the entire sample irrespective of missingness.

3.2.3 Mean imputation

In mean imputation each missing value for a given variable is replaced (imputed) by the observed mean for that variable. Mean imputation is well known to produce biased coefficient estimates in linear regression models even when missingness is MCAR (Little 1992). Standard errors also tend to be too small, giving confidence intervals that are too narrow or tests that reject the null hypothesis more frequently than the nominal value would suggest.

3.2.4 Mean imputation with a dummy

Mean imputation with a dummy is a simple extension of mean imputation (Anderson et al. 1983). Missingness is imputed by the observed mean value for the variable with missing data, but now the covariate list of the (generalized) regression is extended to include a dummy variable $D = 1$ if a case is missing on some X, and $D = 0$ otherwise. If there are several variables with missing observations, then a dummy variable corresponding to missingness on each of these variables is included in the (generalized) regression. This is a common approach to missingness in multivariate regression analyses, because the missingness dummy can be used as a diagnostic tool for testing the hypothesis that the missing data are missing completely at random: If the dummy coefficient is significant, then the data are not MCAR.

Mean imputation with a dummy has properties similar to those for mean imputation without a dummy. Even with the dummy, coefficient estimates can still be biased (Jones 1996).

3.2.5 Conditional mean imputation

In conditional mean imputation, missing values for some variable X are replaced by means of X conditional on other variables in the data set. Typically these means are the predicted values from a regression of X on other covariates in the substantive model, although this restriction is not required. However, if Y is included, results will be biased because of "over fitting" (Little 1992). We shall return to this point in the discussion of the approximate Bayesian bootstrap and Bayesian multiple imputation, both of which use Y in the imputation model.

For data on which conditional mean imputation has been used, linear regression coefficients in the substantive model are biased but consistent (Little 1992). Estimated

substantive models in which missing values have been filled in by conditional mean imputation will tend to under-estimate the standard errors of the regression coefficients, because the standard errors do not account for uncertainty in the imputed values.

3.2.6 Hotdeck imputation

Hotdeck imputation (Brick and Kalton 1996) uses a random draw from an imputation class to fill in each missing datum. Within each imputation class a missing observation on X is replaced by randomly sampling a single observed value of X (with replacement) from that class. Imputation classes for hotdecking are analogous to the weighting classes discussed for weighted casewise deletion.

The number of imputation classes is typically kept small for tractability. Too few classes will result in coefficient bias in the substantive model. Too many classes will increase coefficient variability. Little and Rubin (2002) suggest that three to five strata will suffice.

When the missingness mechanism is MCAR or MAR and the imputation model is correctly specified—the imputation classes are based on all of the observed data for variables that correlate with X—hotdecking is thought to yield unbiased coefficient estimates.[8] However, because only a single draw is made for a given individual missing on X, hotdecking under the stated condition is statistically inefficient.

Again, as with the other techniques discussed in previous sections, analyzing the completed data (observed and imputed) with standard software will result in biased estimates of standard errors because the estimates do not take into account that the imputed data are a resample of the observed data rather than independently observed.[9]

3.2.7 Multiple imputation

The purpose of multiple imputations of each missing datum is to incorporate variability due to the imputation process into assessments of the precision with which the coefficients of the substantive model are estimated (Rubin 1987). The technique requires that the missing observations be imputed M times (Rubin 1996) indicates that $M = 3$ or $M = 5$ is often sufficient; Royston (2004) suggests the use of a larger M.[10] This creates M imputed data sets, each with a potentially different value for each missing datum on each case with missing data. Using these M data sets, the analyst estimates the substantive model M times, once with each data set. The final estimate for the kth of K regression coefficients in the substantive model is the average of that coefficient over the M regressions (Rubin 1987). The estimated standard error of that coefficient,

[8] Maximum likelihood estimation of a logistic regression model is nearly unbiased even when the data are fully observed (McCullagh and Nelder 1989, pp. 455–456). The claim is that under the asserted condition hotdecking does not contribute further bias.

[9] Rao and Shao (1992) propose a variance correction for single stochastic imputation of a mean. We experimented with a generalization of this technique to logistic regression. While its complexity and difficulty of implementation place it beyond the scope of this article, we found that it increased variance estimates to the expected order.

[10] Rubin offers little in the way of justification for the sufficiency of $M = 3$ to 5; in this article, we have used $M = 10$ in all cases.

however, is not just the average of the standard errors from the M regressions. The standard error estimate combines the within-replicate uncertainty (averaged across the M regressions) with the between-replicate uncertainty (the difference across the M regressions). More specifically, for $m = 1, \ldots, M$, the standard error of a coefficient is obtained using

$$\mathrm{SE}(b) = \sqrt{\sum \frac{\mathrm{SE}^2(b_m)}{M} + \left(\frac{M+1}{M}\right) \sum \frac{(b_m - \bar{b})^2}{M-1}}.$$

Simply averaging over the M estimates of a coefficient in the substantive model and plugging replications into the above formula for coefficient standard errors does not necessarily yield estimates with desirable properties. Much depends on how the researcher imputes M times. A sufficient condition for unbiasedness is that the imputations be "proper" (Rubin 1987, pp. 116–132). If they are, then the coefficients averaged over the M imputations are unbiased and the above variance formula is accurate.

Full Bayesian imputation

Rubin (1987) develops a full Bayesian statistical model for making proper imputations. There are various ways to carry out multiple imputation. Schafer (1997a) provides a general approach to the computation of imputed values. To apply the multiple imputation technique to the LAMP data, we used Schafer's (1997b) *S-Plus*/*R* function. For the simulations to be discussed later, we wrote our own Stata code to specialize Schafer's algorithm to our problem.

Briefly, here is what Schafer's algorithm for discrete data did with the LAMP data. First, it fit a saturated (fully interacted) log linear model based on *all* of the substantive model variables (including Y). Using this model to specify the likelihood and minimally conjugate priors, the function explored the posterior distribution of the missing data using data augmentation (Tanner and Wong 1987; Schafer 1997a). This procedure iterates between parameters and missing data imputations. Specifically, in one cycle of the iterative procedure it produces random draws from the posterior distribution of the parameters and then, conditional on these parameter draws, produces draws for the missing values. Each cycle depends on the updated data that were the result of the last step of the preceding cycle.

We captured the draws of the missing data at every 100th iteration up to the 1,000th iteration. That is, we saved 10 imputations.

Approximate Bayesian bootstrap

Full Bayesian multiple imputation is computationally intensive. The approximate Bayesian bootstrap (ABB) is much less so, and can also provide proper multiple imputations (Rubin 1987; Rubin and Schenker 1986). In ABB imputation, M bootstrap samples of the nonmissing cases are created. A bootstrap sample is a random sample drawn from the original sample with replacement that has the same number of observations as the full data set (Efron and Tibshirani 1993). In ABB, the imputation model is estimated for each bootstrap sample, and missing values in the mth sample are imputed on the basis of the model estimates for that sample. Clearly, the coefficients

of the imputation model will vary slightly over the M bootstrap samples. Rubin and Schenker (1986) show that under some conditions if the imputation model is "good" and includes Y, then ABB imputations are proper. More generally, we expect that ABB will produce better estimates of coefficient standard errors in the substantive model than techniques that make no attempt to account for sampling variability in the imputation model, but cannot be certain that ABB is always fully proper.

4 Application of missingness techniques to the LAMP data

This section presents the results of applying the eight missingness techniques we have described to the LAMP data. Table 1 presents eight versions of a logistic regression of mammography compliance using the LAMP data. The regressions are identically specified, but each is based on a different missingness technique. No perusal of these regressions can reveal or verify the properties of the different techniques. The data are real; we do not know with certainty whether the missingness mechanism is MCAR, MAR, or MNAR; we do not know the true imputation model; nor are we certain that the substantive model is perfectly specified. The exercise is nonetheless of value for two reasons. First, it enables us to ask whether the choice of missingness technique matters with a genuine data set that has been used for policy research. Second, the exercise reveals important features of the data that can be used to construct simulation exercises that are firmly rooted in reality.

For the LAMP data, several conclusions are apparent:

1. How missing data are treated affects results: In regressions 1–2, for case and weighted case deletion, the coefficients for doctor's race/ethnicity and respondent's education and marital status are not significant. In the regressions based on the other missingness techniques, these coefficients are significant.[11]
2. The coefficient for dichotomized household income, the sole variable with missingness, is not significant in any regression. However, this coefficient is similar across regressions 5–8, which use conditioned imputation.
3. When household income is mean imputed (regressions 3–4), its coefficients are smaller, which suggests attenuation.
4. All of the techniques that impute missing data (regressions 3–8) produce similar coefficients and standard errors except for household income and the intercept.

The results presented in Table 1 will not support the conclusion that any missingness technique has performed better than another, since we do not know the "true" parameters of the population from which the LAMP data were drawn. Allison (2001, p. 7) suggests that case deletion may outperform multiple imputation techniques when missingness is MNAR. In an attempt to determine if that is so in the present case, and also to consider related questions, we turn next to simulations based on the LAMP data.

[11] A critic noted that if confidence intervals are drawn for all coefficients in Table 1, virtually all point estimates are contained by the corresponding confidence intervals for each missing data technique. The key point, however, is that these results represent different treatments of a **single draw** from the population, not the hypothetical re-sampling to which confidence intervals refer. That the point estimates differ at all suggests the need for further analysis; thus, our simulations.

Table 1 Logistic regression of mammography compliance status under various treatments of missing data for household income

	(1) Case deletion	(2) Weighted case deletion	(3) Mean imputation	(4) Mean imputation with dummy	(5) Conditional mean imputation	(6) Hot deck	(7) Bayesian bootstrap	(8) Bayesian
Constant	−0.49	−0.50	−0.38	−0.38	−0.49	−0.49	−0.38	−0.38
	(−1.92)	(−1.94)	(−1.49)	(−1.48)	(−1.99)	(−2.07)	(−1.55)	(−1.52)
MD Enthusiasm	0.77	0.71	0.89	0.90	0.89	0.88	0.89	0.88
	(4.16)	(3.23)	(6.17)	(6.22)	(6.08)	(5.97)	(6.12)	(6.05)
HH Income> $10,000	0.26	0.39	0.15	0.14	0.29	0.29	0.26	0.35
	(1.22)	(1.87)	(0.68)	(0.67)	(1.36)	(1.55)	(1.27)	(1.81)
High school graduate	0.31	0.32	0.53	0.54	0.48	0.50	0.51	0.49
	(1.25)	(1.18)	(2.39)	(2.43)	(2.12)	(2.25)	(2.25)	(2.14)
MD Asian[a]	−0.26	−0.31	−0.37	−0.37	−0.38	−0.37	−0.38	−0.38
	(−1.32)	(−1.69)	(−2.17)	(−2.16)	(−2.20)	(−2.17)	(−2.18)	(−2.18)
MD Hispanic[a]	−0.22	−0.28	−0.56	−0.57	−0.56	−0.57	−0.57	−0.57
	(−0.71)	(−0.81)	(−2.56)	(−2.57)	(−2.53)	(−2.53)	(−2.55)	(−2.52)
Same MD 1+ years	0.58	0.49	0.49	0.49	0.49	0.49	0.49	0.49
	(3.14)	(2.53)	(2.66)	(2.66)	(2.62)	(2.63)	(2.63)	(2.62)
Married	0.22	0.25	0.30	0.30	0.28	0.29	0.29	0.27
	(1.60)	(1.73)	(2.54)	(2.54)	(2.39)	(2.46)	(2.39)	(2.29)

Table 1 continued

	(1) Case deletion	(2) Weighted case deletion	(3) Mean imputation	(4) Mean imputation with dummy	(5) Conditional mean imputation	(6) Hot deck	(7) Bayesian bootstrap	(8) Bayesian
Medical insurance	1.19	0.77	0.90	0.90	0.88	0.88	0.88	0.88
	(3.88)	(2.57)	(2.95)	(2.95)	(2.91)	(2.88)	(2.88)	(2.88)
Hispanic	−0.27	−0.32	−0.45	−0.47	−0.42	−0.42	−0.42	−0.41
	(−0.94)	(−1.11)	(−1.69)	(−1.71)	(−1.54)	(−1.53)	(−1.55)	(−1.49)
Missingness dummy				0.05				
				(0.36)				
N	857	857	1,119	1,119	1,119	1,119	1,119	1,119

The response variable is defined as $Y = 1$ if the respondent is in compliance, = 0 otherwise. Numbers in parentheses are ratios of coefficients to standard errors estimated using the sandwich estimator modified to take into account the clustered sampling design. Where multiple imputation is used, application of the modified sandwich estimator takes place separately for each imputed data set

Source: Los Angeles Mammography Promotion in Churches Program, baseline survey

[a] The reference category is "MD of other race/ethnicity"

5 Simulations

This article reports simulations based on the LAMP data; in that sense the simulations are realistic. We generated simulated samples in order to study the performance of missingness techniques in samples where the enthusiasm with which the respondent's doctor supported mammography screening was the variable subject to missingness. We chose to simulate missing doctor enthusiasm rather than missing income because the relationship between the outcome (compliance) and physician enthusiasm for mammography is much stronger than the relationship between compliance and income. We were concerned that simulations based on the weaker relationship might produce less conclusive or even spurious results. We do, however, use the observed frequency of missing incomes as a realistic baseline from which to assign missing enthusiasms.

To generate a "population" that is similar to the LAMP data, we began with the 1,119 observations in the LAMP data set that are complete except for household income. For the 262 cases missing on household income, we imputed using a procedure analogous to the procedure used in ABB (Sect. 3.2.7.2). The originally nonmissing cases, together with the cases for which household income was imputed, constitute the population for the simulation exercise.

We generated 2,500 fully observed bootstrap samples from the population defined above. We treat each bootstrap sample as though it is a simple random sample from the population. For each of the 2,500 fully observed bootstrap samples, we created five samples with 262 cases of missingness on physician enthusiasm for a random subsample of observations. The five samples correspond to different missingness mechanisms: missing completely at random (MCAR); missing at random (MAR); missing not at random (MNAR) with probability of nonresponse weakly related to physician enthusiasm; MNAR with probability of nonresponse moderately related to physician enthusiasm; and MNAR with the probability of nonresponse strongly related to physician enthusiasm. Appendix I (available as a web appendix at http://www.pwp.ccpr.ucla.edu) supplies further details on the realizations of the missingness mechanisms in the data sets. In essence we used a balanced design to which, for a given sample and missingness mechanism, we applied eight missingness techniques. Based on findings from our first simulation run, we went back and added a ninth missingness technique, discussed below. For each of the missingness technique by missingness mechanism combinations we estimated the substantive model for mammography compliance using logistic regression.

When missingness is MAR, the imputation regression model (or imputation classes) in the simulations always includes the variable used to create missing data (whether a respondent is Hispanic), as well as other variables. In this sense the imputation models are comparable, although not identical, across missingness techniques. The same point holds for the nonignorability cases, when doctor enthusiasm as well as whether a respondent is Hispanic is used to create missing data.

Figure 1 summarizes results based on the 112,500 ($5 \times 9 \times 2{,}500$) regressions in terms of absolute bias, where bias is defined relative to the *complete data sample* for each iteration (what you would have found had there been no missingness in your

sample), and not the "population" without sampling.[12] The first column summarizes the performance of each missingness technique for each missingness mechanism. The entries in column one are defined as averaged percent bias over all of the coefficients in the regression. Because bias can be positive for one coefficient and negative for another, we use the absolute value of the percent bias for each coefficient and present the mean over all coefficients.

Specifically, let b_{pt} denote the estimated coefficient for the pth of P covariates ($P = 10$) in the logistic regression fit to the tth of T bootstrap samples ($T = 2,500$) of the *fully observed data*. For the jth missing data mechanism and the kth missing data estimation technique, let b_{ptjk} denote the estimate of the pth coefficient of the substantive model fit to the t_jth subsample with missingness (there are four such subsamples for the tth bootstrap sample) using the kth estimation technique. The percent bias for the coefficient of the pth covariate is then

$$BB_{pjk} = 100\frac{\sum_t (b_{ptjk} - b_{pt})}{\sum_t b_{pt}},$$

where BB_{pjk} is the coefficient bias for a specific covariate normed as a percentage. The BB_{pjk} are calculated for each covariate and their absolute values are averaged over all P. Thus, the entries in column one are

$$\overline{BB}_{jk} = (1/P)\sum\nolimits_p \left|BB_{pjk}\right|. \tag{1}$$

Column two of Fig. 1 displays the percent bias in coefficient standard error estimates. Because the application of most missingness techniques leads to standard error estimates that are too small, we have defined percent bias in the standard errors so that more extreme under-estimation will result in a larger positive percent bias.

For the pth covariate, jth missingness mechanism and kth estimation technique, let s_{pjk} denote the standard deviation of the b_{ptjk}. That is,

$$s_{pjk} = \sqrt{\frac{\sum_t (b_{ptjk} - \bar{b}_{p.jk})^2}{T-1}}.$$

Let se_{ptjk} denote the estimated standard error for the pth coefficient from the logistic regression fit to the tth bootstrap sample subjected to the jth missingness mechanism, using the kth missingness technique. In other words, se_{ptjk} is the usual standard error based on the information matrix of the regression for a given data set. The percent bias in standard error estimates for the pth covariate is

$$BSE_{pjk} = 100\frac{s_{pjk} - \overline{se}_{p.jk}}{s_{pjk}},$$

[12] Because we cannot be certain that the substantive model is perfectly specified, both bias due to specification error and bias due to missingness technique may be present in these results.

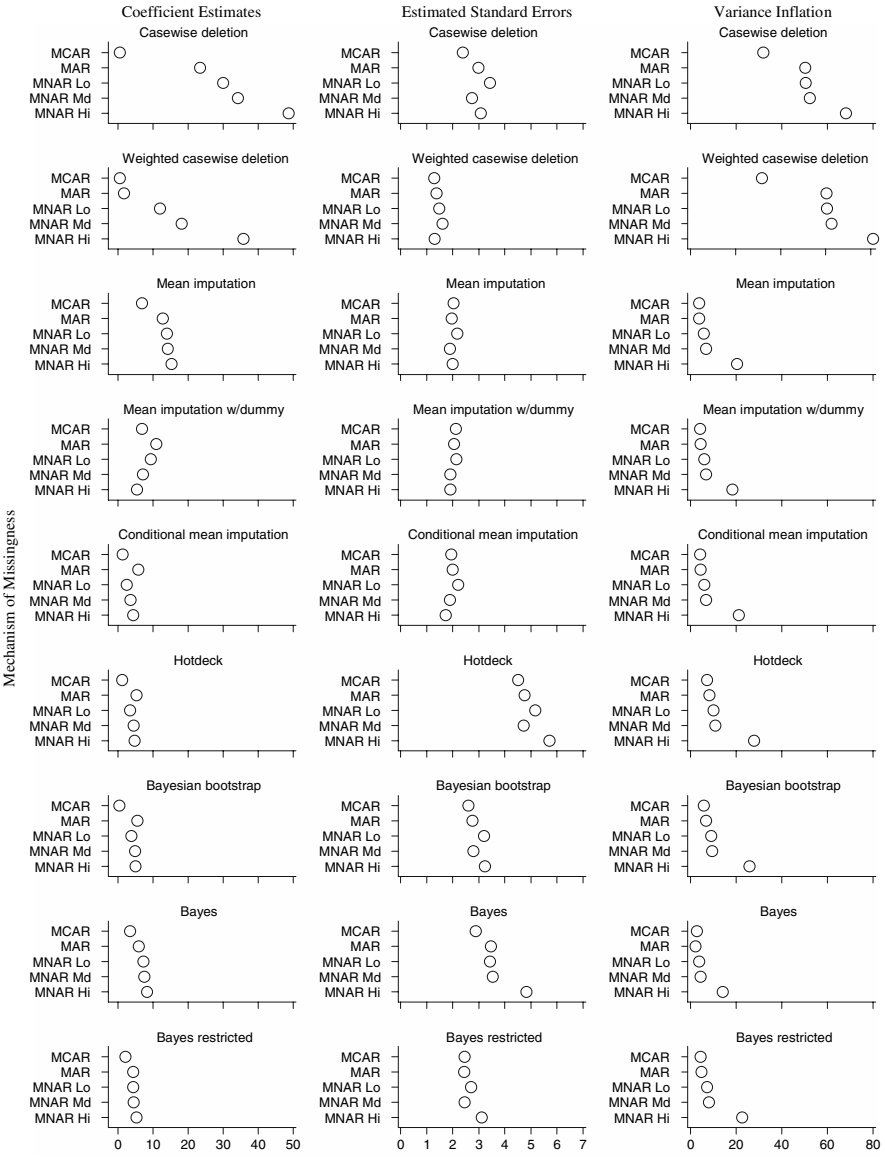

Fig. 1 Observed Bias of Estimates from Simulation of Missing MD Enthusiasm ($N = 2,500$)

where $\overline{se}_{p.jk}$ is the average of the estimated standard errors over the T replicates. The entries in column two of Fig. 1 are then defined to be

$$\overline{BSE}_{.jk} = (1/P) \sum\nolimits_{p} \left| \mathrm{BSE}_{pjk} \right|. \tag{2}$$

Column three in Fig. 1 displays inflation in the variance of the coefficient estimates due to missingness, as a function of the missingness mechanism and the missingness

technique. Let s_p denote the standard deviation of the b_{pt}, the estimate of the pth coefficient in the tth bootstrap sample of the complete data, that is,

$$s_p = \sqrt{\frac{\sum_t (b_{pt} - \bar{b}_{p.})^2}{P-1}}.$$

The percent inflation of variance for the pth coefficient and the jkth combination of missing data mechanism and missing data technique is defined as

$$VI_{pjk} = 100\frac{s^2_{pjk} - s^2_p}{s^2_p}.$$

Large values of VI_{pjk} indicate that missing data results in substantially more variable parameter estimates, conditional on a given combination of mechanism and technique. The entries in column three of Fig. 1 contain the average of the absolute values of the VI_{pjk} over all the coefficients in the substantive model for a particular jk combination:

$$\overline{VI}_{.jk} = (1/P)\sum_p \left|VI_{pjk}\right|. \tag{3}$$

In these simulations, casewise deletion yields unbiased coefficients only in the MCAR case. For other missingness mechanisms, casewise deletion is a poor performer using the criterion of coefficient bias. For all of the missingness mechanisms, casewise deletion is inefficient. Weighted casewise deletion, however, shows virtually no coefficient bias in the MCAR and MAR cases, but is also inefficient.

With respect to coefficient bias, the major divide for these simulations is between the case deletion and imputation techniques. For all of the imputation techniques, coefficient bias is less than for the case deletion techniques as nonignorability increases. All of the imputation techniques have lower variance inflation, because the case deletion techniques are based on fewer observations.

Among the imputation techniques, mean imputation with a dummy performs well under most conditions. It should not (Jones 1996). That it does is more a reflection of the particular details of our simulation setup than it is an indication of the inherent advantages of mean imputation with a dummy. Specifically, the result is a consequence of treating X as a binary coded covariate in conjunction with the way we induced MNAR in the missingness on X. As described in Appendix I, we increased the degree of nonignorability across the three MNAR conditions (low, medium, high) by increasing the odds ratio between X and D (the missingness dummy). Doing so increases the concentration of missingness at $X = 0$. Since X is binary, the impact of the way in which we induce increased nonignorability is to impute with increasing accuracy the typical missing value of X as nonignorability increases.[13] Presumably the result observed here for mean imputation with a dummy will occur in other situations in

[13] This conclusion hinges on the concentration of missingness at a single value of X. The particular imputed value of "mean imputation with a dummy" is irrelevant. Any value will do; inclusion of the missingness dummy in the substantive regression will compensate for error.

which missingness is concentrated at a single value of X, regardless of whether X is discrete or numerically scaled.

Among the imputation techniques, conditional mean imputation has coefficient bias on par with that of hotdecking and the approximate Bayesian bootstrap, and slightly better than that of full Bayesian multiple imputation. Its standard error bias is also modest and stable across missingness mechanisms, and its variance inflation is also quite modest. Hotdecking also performs well, and although its standard error bias is the largest found for the imputation techniques, the levels of bias are modest.

Of the two multiple imputation techniques, the approximate Bayesian bootstrap is essentially on par with full Bayesian imputation, with the ABB coefficients showing less coefficient bias. Both perform well in absolute terms, but with respect to coefficient bias full Bayes does not exceed the performance of conditional mean imputation or hotdecking.

6 Discussion of doctor enthusiasm simulations

6.1 Casewise and weighted casewise deletion

Casewise deletion performs as expected for the MCAR case: There is little or no coefficient bias, the coefficient standard errors are large, and variances are inflated. Under the MAR condition, however, casewise deletion performs poorly—worse than any other technique considered. Allison (2001) suggests that under widely occurring MAR conditions casewise deletion will perform well. Why isn't that performance realized here?

If we were dealing with a linear model problem, it would follow that either the MAR mechanism required an association between missingness on X and values of Y, or that the substantive model was not perfectly specified (Jones 1996). Our setup involves logistic regression. To provide at least partial evidence in support of intuition, we ran simulations to determine if the same conclusions would hold for the logistic regression case. In these simulations, we found virtually no difference in conclusion between results for ordinary least squares and those for logistic regression. From this we infer, since we controlled the MAR mechanism and it did *not* depend on Y, that the substantive model was imperfectly specified.

It will not have escaped notice that we could not have reached this conclusion with the original data; the simulations were essential. The substantive model was plausible and arrived at through reasoned consideration and data analysis. It passed a test of peer review. This substantive model is hardly exceptional, and seems as well specified as many.

Our sense of the "fragility" of casewise deletion is reinforced by further simulations we carried out in response to the casewise deletion results summarized in Fig. 1. Specifically, we ran simulations based on substantive models that included a variety of interactions involving doctor enthusiasm by length of relationship with the respondent; doctor enthusiasm by doctor ethnicity; and a number of interactions with respondent ethnicity. Even with these interactions, the coefficient bias for case deletion in the MAR case was virtually identical with that seen originally. Note also that in the original data, using case deletion, we find no evidence for any of these interactions.

Weighted casewise deletion performs well in the MCAR and MAR cases, with respect to coefficient bias, as it was expected to. Recall that the MAR mechanism is not directly conditioned on Y. Thus, the difference in performance of casewise and weighted casewise deletion can not be due to the broader range of conditions under which weighted case deletion will yield unbiased estimates in the substantive model. Rather, the results summarized in Fig. 1 lead to the inference that the weights correctly capture the components of the missingness model.

With increasing MNAR, the performance of weighted casewise deletion deteriorates. This should not be surprising. The weights are less able to capture the distribution of the observations without missingness because the weights increasingly diverge from the missingness model as X plays an increasingly important role in determining missingness, or in other words, as the factors determining missingness are increasingly located outside the data.

6.2 Imputation techniques

The performance of most imputation techniques when the missingness mechanism is MCAR, MAR, or moderately MNAR is unsurprising. Attempts to impute missing values based on an assumed missingness mechanism that is incorrect in a given instance should result in coefficient bias. Similarly, when an imputation model does not perfectly capture the missingness mechanism but is close, there should be some residual coefficient bias.

We were unprepared for the performance of imputation methods when missingness is MNAR, let alone when it is highly MNAR. Theoretically, none of these techniques is appropriate when missingness is MNAR. To be sure, with the exception of mean imputation with a dummy (explained above), coefficient bias increases with increasing MNAR. Yet, for *all* of the imputation techniques considered here that attempt to model missingness (conditional mean imputation, hot decking, Bayesian bootstrap, and full Bayesian multiple imputation), coefficient bias increases only a small amount as missingness becomes increasingly MNAR, relative to bias under the MAR condition. Also unexpected is the performance of full Bayesian multiple imputation. With respect to coefficient and standard error bias, this technique performed no better than the other imputation techniques. We expected the simulations of full Bayesian multiple imputation to demonstrate virtually unbiased coefficients and standard errors under the MCAR and MAR conditions. What went wrong?

It turns out that the Bayesian multiple imputation model we implemented based on Schafer's (1997a) algorithm is vulnerable to a problem known as "semi-complete separability." To function, the imputation scheme iteratively forms tables of frequencies of $X|\{Y \times Z\}$, where there are, for our setup, four possible outcomes for X—high income, low income, missing, or empty. The empty outcomes do not pose a problem; they simply correspond to $\{Y \times Z\}$ combinations for which there are no data. The problem arises for $\{Y \times Z\}$ combinations for which there are data on X, and for which the data consist of at least one respondent with missingness as well as at least one respondent with no missingness, but for whom all instances are the same on X (either all high income or all low income in this instance) for a given $\{Y \times Z\}$ combination.

In the algorithm we used, it is necessary to iteratively compute probabilities for missing elements based on the observed frequencies of X in a given cell. However, in a semi-completely separated cell, the observed cell proportion is zero for one of the two possible values for any missing elements. The algorithm chooses a nonzero value for that estimated probability based on the cell size and on the minimal conjugate prior specified in the model. In our simulations, that specification results in consistent overestimation of the probability of the unobserved available value appearing in a missing element. If this happened once or twice the consequences would be minimal. However, the low density of cases over the entire set of covariate combinations results in so many semi-completely separated cells that the bias becomes noticeable and considerable.

This was not an obvious problem to us. We expected the simulations to demonstrate the general superiority of Bayesian multiple imputation. Before considering alternative explanations of the bias, we scrutinized our code, certain that the problem must be due to our error. Once we realized semi-complete separability might be a problem, we counted cells in the matrix that were vulnerable to separability as well as cells that actually had a separability event during an iteration of the full Bayesian process. In the LAMP "population" data for the simulation of missing physician enthusiasm, 103 of the 512 cells required for computation of the fully interacted log-linear regression used in the full Bayes model are "possibly" separable (that is, they contain only one of the two binary possibilities for that outcome) and will become separability events if they are "dealt" missing data. Under MCAR, an average of 35 of those cells become separability events in a single iteration.

Had this not been a simulation study in which we had access to the "true" coefficients, *we would never have suspected a problem*. Further, the solution to the problem is not obvious. There would seem to be two possible strategies. First, one could specify a different set of priors that would not induce bias in semi-completely separated cells. While attractive, this is hard to do in practice, and would require unimaginable personal knowledge of the data and the mechanism of missingness, and would certainly preclude use of a generic "black box" multiple imputation algorithm such as the one we used. The second strategy is to eliminate the occurrence of semi-complete separability. This could be accomplished by reducing the complexity of the model (Schafer 1997a, p. 341). For example, had we been imputing from a fully interacted log-linear model based on $X|\{Y \times Z^*\}$, where Z^* is a judiciously chosen subset of Z, we probably could have avoided semi-complete separability, but to do so *we might have had to use an imputation procedure that failed to include all of the covariates in the correct (i.e., " true") imputation regression model.* We will refer to this strategy as "restricted" Bayes imputation.

To put the problem another way: We had what was thought to be the "correct" imputation model, but it "over-taxed" the data. To stay within the limits of the data, we would have to reduce the imputation model, which might mean that we no longer had the correct model. We would not know whether we did. Further, we would have no obvious means to discern how far from "correct" our reduced model was.

It is true that if the researcher has a single, real data set, it is possible to observe whether the imputation process is encountering semi-complete separability. It did not occur to us to check for this possibility when using our own code until we studied the simulation results, and Schafer's code does not provide the necessary window.

Researchers intent on multiple imputation are advised to be aware of the need to check for semi-complete separability. But even if checking is done, how much semi-complete separability is too much? Further, if one decides to simplify the imputation model to eliminate semi-complete separability, then it is necessary to enter the realm of model uncertainty. Here the question is, how does one know when the imputation regression specification is "good enough?" As a step toward answering this question, we applied a restricted Bayes imputation technique to the simulations of missing physician enthusiasm (see the final row of Fig. 1). To obtain a model with no chance of separability we used a fully interacted log linear specification that contained only four variables, instead of the nine that are used in the full model. The four variables that avoided separability in their 32 cell interaction space were: compliance (Y), married or living with a partner, patient ethnicity, and physician ethnicity.

Figure 1 shows that, with respect to coefficient bias, the restricted Bayes specification performs better than full Bayes, and about the same as conditional mean imputation, hotdecking, and the approximate Bayesian bootstrap. Resolution of the semi-complete separability problem in the chosen way seems to improve the performance of Bayesian multiple imputation.

7 Conclusion

Based on our simulation analysis we find that casewise deletion is particularly vulnerable to imperfections in the substantive model. What might those imperfections be? The additive logistic regression for mammography screening compliance used in the work reported in this article was arrived at by Fox et al. using a data set that was designed for the analysis reported by Fox et al. (1998). We re-checked the data for a number of interactions using a substantively based search procedure rather than an automated interaction detector, and found no evidence of interaction. In this respect our data analysis was at least conventional and might possibly be viewed by some as careful. We could have missed some substantively plausible interactions; if so, then there will be others who will do likewise in the analysis of their own data. If in fact there are no relevant interactions between the covariates we considered for the study of mammography screening compliance, this would seem to leave three possibilities: covariate measurement error, omitted covariates, and inappropriate application of the logistic distribution. We will not speculate about these possibilities except to note that (i) the reduction of the problem to one that is expressed entirely in terms of discrete variables, often simplified to dichotomies, is defensible; (ii) although there may be omitted covariates, the data were collected specifically to study the problem of mammography screening compliance, and the regression includes a number of appropriate covariates; (iii) the application of logistic regression to a substantive problem for which the observed variables are fully discrete would generally be considered appropriate. For these reasons we conclude that even in cases where missingness is known to be MAR and not to depend on Y, aspects of the estimated substantive model that might not be viewed as imperfections by substantive researchers apparently can result in considerable coefficient bias when casewise deletion is used.

If the analyst is able to arrive at a defensible imputation model based on other variables and using conditional mean imputation; hot decking; the approximate Bayesian

bootstrap; or full Bayesian multiple imputation, it is possible to obtain results with mild coefficient bias—even, surprisingly, when missingness is somewhat MNAR. Unfortunately, there is a caveat, which is that there is a need for a reasonably well-specified model of missingness, and a similarly well-specified imputation model (Landerman et al. 1997). The imputation of missing values is a substantive data analysis problem deserving no less attention than the substantive problems that attract analysts to data in the first place. Misspecification of the imputation model can result in a degree of coefficient bias that is as bad as, or worse than, that produced by case deletion.

The problem of possible imputation model misspecification is not easily solved. Full Bayesian multiple imputation, a technique that purports to take care of imputation modeling for the analyst by making use of all available covariates and their relationships, can exceed the limits of the data by creating a situation with extensive semi-complete separability. If the imputation model is constrained so that it does not over-tax the data, there is then a risk that it is incorrect. Thus, if the separability problem occurs and the analyst is aware of it, there nonetheless remains uncertainty about the impact of the specification of the imputation model on coefficient bias in the substantive model. This is true for any missingness technique, but in the simulations we examined, it is especially so for Bayesian multiple imputation.

In the final analysis, we conclude that: (1) Automated selection of the imputation model in the use of full Bayesian multiple imputation can lead to greater bias in coefficients of substantive models than that resulting from other missing data techniques we tried. (2) Under conditions that occur in actual data, casewise deletion can perform less well than we were led to expect by the existing literature. (3) Relatively unsophisticated imputations, such as mean imputation and conditional mean imputation, performed better than the technical literature led us to expect. (4) Imputation models *are* substantive models, and require the same caution with respect to specificity and calculability.

References

Allison PD (2001) Missing data. Sage Publications, Thousand Oaks

Ambler G, Omar RZ (2007) A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. Stat Methods Med Res 16:277–298

Anderson AB, Basilevsky A, Hum DPJ (1983) Missing data: a review of the literature. In: Rossi, Wright, Anderson (eds) Handbook of survey research. Academic Press, New York

Breen N, Kessler L (1994) Changes in the use of screening mammography: evidence from the 1987 and 1990 National Health Interview Surveys. Am J Public Health 84:62–72

Brick JM, Kalton G (1996) Handling missing data in survey research. Stat Methods Med Res 5:215–238

Carpenter JR, Kenward MG, White IR (2007) Sensitivity analysis after multiple imputation under missing at random: a weighting approach. Stat Methods Med Res 16:259–275

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York

Farewell VT (1979) Some results on the estimation of logistic models based on retrospective data. Biometrika 66:533–538

Fox J (1997) Applied regression analysis, linear models, and related methods. Sage Publications, Thousand Oaks

Fox SA, Siu AL, Stein JA (1994) The importance of physician communication on breast-cancer screening of older women. Arch Intern Med 154:2058–2068

Fox SA, Pitkin K, Paul C, Carson S, Duan N (1998) Breast cancer screening adherence: does church attendance matter? Health Educ Behav 25:742–758

Groves RM, Singer E, Corning A (2000) Leverage–Saliency theory of survey participation. Public Opin Q 64:299–308
Heckman J (1976) The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models. Ann Econ Soc Meas 5:475–492
Heckman J (1979) Sample selection bias as a specification error. Econometrica 47:153–161
Jones MP (1996) Indicator and stratification methods for missing explanatory variables in multiple linear regression. J Am Stat Assoc 91:222–230
Landerman LR, Land KC, Pieper CF (1997) An empirical evaluation of the predictive mean matching method for imputing missing values. Sociol Methods Res 26:3–33
Little RJA (1992) Regression with missing X's: a review. J Am Stat Assoc 87:1227–1238
Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, New York
McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall, New York
Rao JNK, Shao J (1992) Jackknife variance estimation with survey data under hot deck imputation. Biometrika 79:811–822
Royston P (2004) Multiple imputation of missing values. Stata J 4:227–241
Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York
Rubin DB (1996) Multiple imputation after 18+ years. J Am Stat Assoc 91:473–489
Rubin DB, Schenker N (1986) Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. J Am Stat Assoc 81:366–374
Rubin DB, Schenker N (1991) Multiple imputation in health-care databases: an overview and some applications. Stat Med 10:585–598
Schafer JL (1997a) Analysis of incomplete multivariate data. Chapman & Hall, London
Schafer JL (1997b) Software for multiple imputation. [http://www.stat.psu.edu/~jls/misoftwa.html]
Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation (with discussion). J Am Stat Assoc 82:528–550
Vach W (1994) Logistic regression with missing values in the covariates. Springer, New York
Xie Y, Manski CF (1989) The logit model and response-based samples. Sociol Methods Res 17:283–302