California Center for Population Research
University of California - Los Angeles

What Should We Do About Missing Data?

(A Case Study Using Logistic Regression with Missing Data on a Single Covariate)*

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

CCPR-028-03

October 2003

California Center for Population Research
On-Line Working Paper Series

What Should We Do About Missing Data?

(A Case Study Using Logistic Regression with Missing Data on a Single Covariate)*

Christopher Paul[a], William M. Mason[b], Daniel McCaffrey[c], and Sarah A. Fox[d]

Revision date: 24 October 2003

File name: miss_pap_final_24oct03.doc

[a] RAND, cpaul@rand.org
[b] Department of Sociology and California Center for Population Research, University of California–Los Angeles, masonwm@ucla.edu
[c] RAND, Daniel_McCaffrey@rand.org
[d] Department of Medicine, Division of General Internal Medicine and Health Services Research, University of California–Los Angeles, sfox@mednet.ucla.edu

*The research reported here was partially supported by National Institutes of Health, National Cancer Institute, R01 CA65879 (SAF). We thank Nicholas Wolfinger, Naihua Duan, and John Adams for comments on an earlier draft.

ABSTRACT

Fox et al. (1998) carried out a logistic regression analysis with discrete covariates in which one of the covariates was missing for a substantial percentage of respondents. The missing data problem was addressed using the "approximate Bayesian bootstrap." We return to this missing data problem to provide a form of case study. Using the Fox et al. (1998) data for expository purposes we carry out a comparative analysis of eight of the most commonly used techniques for dealing with missing data. We then report on two sets of simulations based on the original data. These suggest, for patterns of missingness we consider realistic, that case deletion and weighted case deletion are inferior techniques, and that common simple alternatives are better. In addition, the simulations do not affirm the theoretical superiority of Bayesian Multiple Imputation. The apparent explanation is that the imputation model, which is the fully saturated interaction model recommended in the literature, was too detailed for the data. This result is cautionary. Even when the analyst of a single body of data is using a missingness technique with desirable theoretical properties, and the missingness mechanism and imputation model are supposedly correctly specified, the technique can still produce biased estimates. This is in addition to the generic problem posed by missing data, which is that usually analysts do not know the missingness mechanism or which among many alternative imputation models is correct.

1. Introduction

The problem of missing data in the sense of item nonresponse is known to most quantitatively oriented social scientists. Although it has long been common to drop cases with missing values on the subset of variables of greatest interest in a given research setting, few data analysts would be able to provide a justification, apart from expediency, for doing so. Indeed, probably most researchers in the social sciences are unaware of the numerous techniques for dealing with missing data that have accumulated over the past 50 years or so, and thus are unaware of reasons for preferring one strategy over another. Influential statistics textbooks used for graduate instruction in the social sciences either do not address the problem of missing data (e.g., Fox 1997) or present limited discussions with little instructional specificity relative to other topics (e.g., Greene 2000). There are good reasons for this. First, the vocabulary, notation, acronyms, implicit understandings, and mathematical level of much of the missing data technical literature combine to form a barrier to understanding by all but professional statisticians and specialists in the development of missing data methodology. Translations are scarce. Second, overwhelming consensus on the one best general method that can be applied to samples of essentially arbitrary size (small as well as large) and complexity has yet to coalesce, and may never do so. Third, easy to use "black box" software that reliably produces technically correct solutions to missing data problems across a broad range of circumstances does not exist.[1]

Whatever the method for dealing with missing data, substantive researchers ("users") demand specific instructions, and the assurance that there are well documented reasons for accepting them, from technical contributors. Absent these, researchers typically revert to case deletion to extract the complete data arrays essential for application and interpretation of most

[1] Horton and Lipsitz (2001) review software for multiple imputation; Allison (2001) lists packages for multiple imputation and maximum likelihood.

multivariate analytic approaches (e.g., multiway cross-tabulations, the generalized linear model). For, despite its potential to undermine conclusions, the missing data problem is far less important to substantive researchers than the research problems that lead to the creation and use of data.

This paper developed from a missing data problem: Twenty-eight percent of responses to a household income question were missing in a survey 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. With that head start, we originally intended only to exposit several of the more frequently employed strategies for dealing with missingness, using the missing household income problem for illustration. Of course, application of different missingness techniques to the same data can not be used to demonstrate the superiority of one technique over another. For this reason as well as others, we then decided to carry out simulations of missing household income, in order to illustrate the superiority of Bayesian stochastic multiple imputation and the approximate Bayesian bootstrap. This, we thought, would stimulate the use of multiple imputation. The simulations, however, did *not* demonstrate the superiority of multiple imputation. In addition, the performance of case deletion was not in accord with our expectations. For reasons that will become clear, we conducted new simulations, again based on the original data. This second round also failed to demonstrate the superiority of multiple imputation, and again the performance of case deletion was not in accord with our expectations.

The source of these discrepancies is known to us only through speculation informed by the *pattern* of performance failures in the simulations. If our interpretation is correct, the promise of these techniques in actual practice may be kept far less frequently than has been supposed. Thus, to the original goal of pedagogical exposition we add that of illustrating pitfalls in the application of missingness techniques that await even the wary.[2]

In Section 2 of this paper we describe the data and core analysis that motivate our study of missingness. Sections 3 and 4 review key points about mechanisms of missingness and techniques for handling the problem. Section 5 presents results based on the application of alternative missing data methods to our data. Section 6 describes the two sets of simulations based on the data. Sections 7 and 8 review and discuss findings. Appendix I contains a technical result. Appendix II details the simulation process. Appendix III provides Stata code for the implementation of the missingness techniques. Upon acceptance for publication, Appendices II and III will be placed on a website, to which the link will be provided in lieu of this statement.

2. Data and Core Analysis

Breast cancer is the most commonly diagnosed cancer of older women. Mammography is the most effective procedure for breast cancer screening and early detection. The National Cancer Institute (NCI) recommends that women aged 40 and over receive screening mammograms every one or two years.[3] Many women do not adhere to this recommendation. To test possible solutions to the under-screening problem, the Los Angeles Mammography

[2] The technical literature on missing data is voluminous. The major monographs are by Little and Rubin (2002), Rubin (1987), and Schafer (1997). Literature reviews include articles by Anderson et al. (1983), Brick and Kalton (1996), and Nordholt (1998). Schafer (1999) and Allison (2001) offer helpful didactic expositions of multiple imputation.

[3] The lower age limit has varied over time. Currently it is age 40. Our data set uses a minimum of age 50, which was in conformance with an earlier guideline.

Promotion in Churches Program (LAMP) began in 1994 (Fox et al. 1998). The study sampled women aged 50-80, all of whom were members of churches selected in a stratified random sample at the church level. In the study, each church was randomly assigned to one of three interventions.[4] The primary analytic outcome, measured at the individual level, was compliance with the NCI mammography screening recommendation. In this study we use data from the baseline survey (N = 1,477), that is, data collected prior to the interventions that were the focus of the LAMP project.[5] Our substantive model concerns the extent and nature of the dependence of mammography screening compliance on characteristics of women and their doctors, prior to LAMP intervention.

In our empirical specification, 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[6] 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

[4] This design, known as “multilevel” in the social sciences, is regarded in biomedical and epidemiological research as an instance of a “group-randomized trial” (Murray 1998).

[5] From a realized sample size of 1,517 individuals we dropped four churches, each with 10 respondents, prior to the analyses reported here and in Fox et al. (1998). This reduced the sample to 1,477 individuals before exclusions due to missingness on any variable in the regression model other than household income. The churches in question were dropped from the LAMP panel due to administrative inefficiencies associated with their small sample size and low participation rates.

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

respondent's doctor is Asian, Hispanic, or belongs to another race/ethnicity group (the reference category in our regressions). Prior research and theory (Breen and Kessler 1994) suggest that those of higher socioeconomic status should be more likely to be in compliance, as should those whose doctors are enthusiastic about mammography, have a regular doctor, are married or have a partner, and have some form of medical insurance. Similarly, there are a priori grounds for expecting women with Asian or Hispanic doctors to be less likely than those with doctors of other races/ethnicities to be in compliance, and for expecting Hispanic women to be less likely than others to be in compliance (Fox et al. 1998; Zambrana et al. 1999).

Deletion of a respondent if information is missing on any variable in the model, including the response variable, 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.[7] 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. We

[7] 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.

choose to focus on this missingness problem, so defined, because of its potential importance for substantive conclusions based on the LAMP study and because restriction of our attention to nonresponse on a single variable holds the promise of greatest clarity in comparisons across techniques for the treatment of missingness.

We suspect that household income was not reported largely because the item was perceived as invasive, not because it was unknown to the respondent. The desire to keep household income private seems likely to be related to income itself or to other measured characteristics—possibly those included in the mammography compliance regression. If so, failure to take into account missingness on household income could not only lead to bias in the household income coefficient but also propagate bias in the coefficients of other covariates in the mammography compliance regression (David et al. 1986). Missingness on household income thus 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 and Models

Three types of models are inherent to all missing data problems: a model of missingness, an imputation model, and a substantive model. A missingness model literally predicts whether an observation is missing. For a single variable with missing data, the missingness model might be a binary (e.g., logistic) regression model in which the response variable is whether or not an observation is missing. This type of model is discussed more precisely in the next section. In that discussion, we also categorize types of missingness models.

An imputation model is a rule, or set of rules, for treatment of missing data. Imputation models can often be expressed as estimable (generalized) regression specifications based on the observed values of variables in the data set. The purpose of such a regression is to produce a value to replace missingness for each missing observation on a given variable.

A substantive model is a model of interest to the research inquiry. In general, our concern is with the nature and extent to which a method for modeling missing data affects the estimated parameters of the substantive model, and with the conditions under which the impact of a method varies.

Missingness models and imputation models do not differ in any meaningful way from substantive models—they are not themselves "substantive" models simply because they are defined relative to a concern with missingness in some other process of greater interest, that is, in some other model. In actual substantive research, researchers generally do not know the correct model of missingness or the correct imputation model (much less the correct substantive model). This lack of knowledge is not a license to ignore missingness. To do so is equivalent to assuming that missingness is completely random, and this can and should be checked. Moreover, the development of missingness and imputation models with reference to a given missing data problem is neither more nor less demanding than the development of the substantive model. From this we conclude: (i) For any substantive research project, missingness and imputation models can and should be developed; (ii) the process of arriving at reasoned missingness and imputation models is no more subject to automation than is the development of the substantive model. Given these models, we ask which techniques excel unambiguously, and whether any achieve a balance of practicality and performance given current technology.

4. 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, and for the extent to which they reshape coefficient distributions to have inaccurate variances (*Var*(b)), where "bias" and "inaccuracy" are specified relative to samples with no missing data. 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.[8] 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 paper is plausible, we do not know if it is perfectly specified, and our simulation analyses reveal that probably it is not.

Table 1 summarizes the received performance of missingness techniques conditional on mechanisms of missingness. The distillation of the technical literature represented by Table 1 assumes that the substantive model is perfectly specified. As can be seen, the technique by mechanism interaction precludes a simple summary. However, the two Bayesian techniques appear to have the best expected performance on the three criteria we have listed.

Insert Table 1 Here

The mechanism assumed to underlie missingness on a particular variable in a given data set ideally has a role to play in broadly determining the type of technique to be used to compensate for the missingness. Our summary in Section 3.1 of the missingness mechanisms used in Table 1 is based on Rubin's typology (Little and Rubin, 2002; Rubin 1987), expressed in

[8] For the case considered in this paper—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.

the development of Bayesian stochastic multiple imputation. Of the eight missingness techniques we consider, six are based on the imputation of missing values.[9] In the case of the LAMP data, imputation means that each respondent who did not supply an answer to the household income question would be assigned one or more estimated values. All of the imputation techniques we consider use the assumption that a substantive model of interest can be estimated independently from—without reference to—both the underlying model for missingness (which might be no more than implicit) and the imputation model. The mechanisms of missingness typology clarifies a *necessary* condition under which missingness is consistent with separation of substantive modeling from missingness and imputation modeling. We next review the mechanisms listed in Table 1, and subsequently describe the techniques.

4.1 Mechanisms of Missingness

All of the missingness or item nonresponse we are concerned with has a random component. In the LAMP survey, women under the age of 50 are excluded by design. Hence all responses of women less than age 50 are necessarily "missing." This nonstochastic missingness is of no interest to us. We begin with this obvious point because the following brief summary of mechanisms of missingness introduces jargon that uses the term "random" in a way not commonly seen elsewhere.

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 (and recodes, combinations, and transformations thereof) in the LAMP data other than Y and X. Mechanisms of missingness can be defined with reference to

[9] We do not consider the "maximum likelihood" technique, largely because it does not appear to be widely used by researchers, and because it does not seem to have received the attention accorded to Bayesian multiple imputation. Allison (2001) provides a helpful introduction to the maximum likelihood technique for missing data; Schafer (1997) and Little and Rubin (2002) provide technical expositions.

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 $R_i = 1$ is independent of X, but not of (some subset of) Y and Z;
3. The probability that $R_i = 1$ depends on X and (some subset of) Y and Z.

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). Equivalently, an appropriate model we construct for predicting R will have only an intercept—all covariates in the prediction model, including the actual values of X (which will be unobserved for some respondents) will have coefficients equal to zero. If missingness is MCAR, then the observed sample yields unbiased estimates of all quantities of interest. The estimates have inflated variance compared to what would be found if there were no missing data.

The second missingness mechanism is known as *m*issing *a*t *r*andom (MAR). Missingness on household income is MAR if it depends on (some subset of) mammography compliance and the remaining variables in the LAMP data, but does not depend on the actual value (even if unobserved) of household income itself once the variables that nonresponse does depend on have been taken into account. Equivalently, in the population from which the LAMP sample has been drawn, there is a value of household income for each potential respondent, some of whom are missing on household income in the sample. Under the MAR assumption, an appropriate prediction model for missingness defined on the population from which the LAMP sample was drawn will have a coefficient equal to zero for household income itself; at least one coefficient for another variable in the LAMP data will not be zero. If missingness is MAR, then the observed sample does not in general yield unbiased estimates of all quantities of interest.

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

That the probability of missingness does not depend on the level of the variable with missingness in the MAR and MCAR cases implies that missingness is independent of variables that are not in the data set. When this major, double-barreled, assumption is 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.[10]

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 by implication, variables not in the data) as well as potentially other variables in the data. 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.[11]

It is in general difficult to know whether missingness is ignorable, especially with cross-sectional data, and it seems a plausible conjecture that some degree of nonignorability in

[10] For other conditions, see Schafer (1997:10).

[11] 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 paper.

missingness processes is common.[12] Here, as in many other situations, a continuum is probably more realistic than an "all or none" typology, and 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.

4.2 Missingness Techniques

4.2.1 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. Called "casewise" or "listwise" deletion, this method is simple to implement. 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 in which Y is not a covariate in the missingness model (equivalently, Y is uncorrelated with missingness on X).[13] If either of these assumptions are satisfied, then unbiased coefficient estimates may be obtained without imputation. Also, the coefficient standard errors will be valid for a sample of reduced size.

Casewise deletion uses less of the available data than the other methods, because observations that are missing on even a single variable (so-called "partially observed records")

[12] Groves et al. (2000) document an instance of nonignorability using a two-wave panel study.

[13] The discussion of theorem 2.1 in Jones (1996) provides the basis for this assertion; see also Allison (2001:6).

are dropped. In addition, it can lead to biased coefficient estimates if any of the above assumptions are violated.

For the LAMP data and our simulation study, casewise deletion on household income reduces the sample size to 857 out of a possible 1119 observations, which is a 23 percent reduction.

4.2.2 Weighted casewise deletion

Weighted casewise deletion extends the range of MAR models under which unbiased coefficient estimation in the substantive model can be achieved.[14] 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. Typically, complete cases are stratified by covariates thought to explain systematic differences between complete and incomplete cases. Within each stratum, the complete cases are given the weight of both the complete and the incomplete cases. For example, in the LAMP data, approximately 56 percent of Hispanic respondents were missing household income, compared with 16 percent of African Americans and 15 percent of non-Hispanic white respondents. Stratifying by race/ethnicity and restricting attention to complete cases, Hispanics would be weighted by 1 + (proportion reporting/proportion missing), which is 1 + .56/.44 = 2.27. African Americans would be weighted by 1.19 and whites by 1.18.

Although weighted casewise deletion can reduce coefficient bias, the technique is inefficient because the exclusion of observed data from partially complete observations reduces

[14] Other names for weighted casewise deletion are casewise re-weighting and nonresponse weighting.

sample size.[15] In addition, unequal weights can increase the variability of the estimates (Cochran 1977).

Care in the application of weights is required if valid standard errors are to be obtained. Fortunately, several software packages provide valid standard errors for nonresponse weighting.[16] Successful application of weighted casewise deletion depends not only on sufficiently accurate and deep substantive knowledge and familiarity with the data but also on satisfying the MAR assumption to some degree.

To apply weighted casewise deletion to the LAMP data, we created 12 weighting classes based on respondent race/ethnicity; health insurance status; and responses to a question concerning general household financial well-being without actual dollar amounts.[17] Cases missing on household income in each weighting class were counted and then dropped. Cases remaining in each weighting class were weighted by the ratio of the total number of cases in the class to the number of cases in the class with household income data, so that the aggregate weight in each class is equal to the total number of cases in each class before deletions. Appendix I, section 2 contains the Stata code we used to implement weighted casewise deletion.

4.2.3 Mean imputation

In mean imputation each missing value for a given variable is replaced (imputed) by the observed mean for that variable. This approach requires only a single calculation (of the mean) and a single data management step (replacement of missing values with that mean). As with casewise deletion, the missingness and imputation models have no covariates, by assumption.

[15] Note, however, that according to Robins et al. (1994), weighted casewise deletion can be fully semi-parametric efficient, which is less than fully model efficient.

[16] For example, SAS proc reg and Stata with analysis weights do not provide valid standard errors with nonresponse weights, although coefficients are correctly estimated. However, Stata's "pweight" option will provide valid standard errors. Cohen (1997) discusses weighting in various statistical packages.

[17] Information about general household financial well-being was provided by a large proportion of those who were unwilling or unable to provide household income.

Mean imputation is well known to produce biased coefficient estimates in linear regression models even when observations are missing completely at random (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.

To apply mean imputation to the LAMP data, for those respondents missing on household income we replaced the missing value code with the mean of the dichotomized household income variable (0.84). Appendix I, section 3 contains the Stata code we used to implement mean imputation.

4.2.4 Mean imputation with a dummy

Mean imputation with a dummy is a simple extension of mean imputation (Anderson et al. 1983). In this method missingness is again 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). Implementation is simple. The technique does, however, leave the analyst with an additional coefficient to interpret for each variable with missingness. The advantage of the technique probably resides in its potential to provide improved predictions. We do not address this aspect of the technique in our simulations.

For the LAMP data we imputed mean household income (0.84) as in mean imputation and included an imputation dummy variable in the list of covariates of the core regression model. Appendix I, section 4 contains the Stata code we used to implement mean imputation with a dummy.

4.2.5 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 imputatio n, both of which use *Y* in the imputation model.

Conditional mean imputation can also be implemented using fully observed covariates to stratify the data into a small number of imputation classes, such as the classes used for casewise reweighting. A missing value on *X* for a given individual is then replaced by the observed conditional mean on *X* for the imputation class to which the individual belongs. Predicted values from a regression will be the same as the observed conditional means of imputation classes when the regression covariates are discrete and fully interacted, and the imputation classes correspond to the cells of the saturated interaction defined by the regression model.[18]

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

[18] If *X* is dichotomous and coded 1 or 0, the imputed values are nonetheless fitted proportions. For a given imputation class, this is equivalent to imputing the correct proportion of 1's and 0's.

substantive model is binary, and logistic regression is used, then the coefficient of the covariate containing imputed values tends to be attenuated regardless of sample size (see Appendix II for the outline of a proof). In addition, 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.

Even in statistical packages that do not specifically implement conditional mean imputation, the technique can be straightforward to implement, requiring only a modeling step and an imputation step prior to "complete case" analysis.[19] For the LAMP data we fit a logistic regression of the dichotomized household income variable on respondent's race/ethnicity, insurance status, general financial well-being (which does not refer to exact dollar amounts), and education. (Apart from education, these covariates were used to create the weighting classes for our weighted casewise deletion analyses.) Since by subsample selection (see Section 2), individuals missing on household income were not missing on the covariates, we then applied the coefficients to the covariate values for these individuals in order to generate predicted household income values. Appendix I, section 5 contains the Stata code we used to implement conditional mean imputation.

4.2.6 Hotdeck imputation

Mean imputation, with or without a dummy, produces a single imputed value that is an estimate of the expected values of the missing observations for a given X. Similarly, conditional mean imputation produces imputed values that are estimates of the expected values of the missing data given the values of observed covariates. If we actually observed any given missing data point it would tend to be close to its imputed value, but not exactly equal to it. Hence,

[19] Stata 8 implements conditional mean imputation via multiple regression in its "impute" command.

imputed values capture only a portion of the variability that would be observed were all the data present. This complete data variability can be captured in the imputed values by using a technique that randomly selects between likely values, or through the addition of random errors to the (conditional) mean imputations. Techniques that introduce a random component to imputation are said to be stochastic. We discuss three: hotdeck; Bayesian multiple imputation; and the approximate Bayesian bootstrap (ABB). Typically, hotdeck imputation uses only a single random imputation for each missing observation. The Bayesian and ABB approaches use multiple random draws to impute multiple possible values for each missing observation.

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 and the strata used for conditional mean imputation.

When macros or dedicated software are not available, the number of imputation classes typically is kept relatively 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 often 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.[20] However, because only a single draw is made for a given individual missing on X, hotdecking under the stated condition is statistically inefficient.

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

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.[21]

Hotdecking is not a standard component of the major statistical packages, although macros are available for several. Most packages have readily employed tools for randomization and internal sampling, which allow for straightforward programming of the technique.

For the LAMP data we performed a single hotdeck draw for each individual missing on household income, using the same 12 imputation classes introduced for the casewise re-weighting example. Appendix I, section 6 contains our Stata code for the implementation of this technique.

4.2.7 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) proposed a technique to do this. The technique requires that the missing observations be imputed M times (Rubin (1996) indicates that $M = 3$ or $M = 5$ often suffices). 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, however, is not just the average of the standard errors from the M models. The standard error estimate combines the

[21] 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 paper, we found that it increased variance estimates to the expected order.

within-replicate uncertainty (averaged across the *M* regressions) with the between-replicate uncertainty (the difference across the *M* regressions). More specifically, for *m = 1,…,M*, the standard error of a coefficient is obtained using

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

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.

The first requirement of proper imputation is that the coefficients of the imputation model must be (nearly) unbiased and consistent, and that the specification of the imputation model must be consistent with the posited mechanism of missingness. In practice, this means (i) that the imputation model must be a "good" model for predicting missingness, and (ii) that if there is any association between the variable with missing data (*X*) and the outcome variable in the substantive model (*Y*), then *Y* must be included in the imputation model.[22]

The second requirement of proper imputation is that it must capture the variability in the estimated parameters of the imputation model. Repeated hotdeck draws, for example, do not constitute "proper" imputation because they do not capture population level uncertainty about the missing data, only sample level uncertainty. A proper imputation model must be structured to

[22] As Allison (2001:53) points out, in Bayesian multiple imputation and the approximate Bayesian bootstrap, the imputed values are not an exact function of *Y* and *Z*. This stochastic aspect of the imputations removes part if not all of the objection to the inclusion of *Y* in the imputation model.

account for the variability in parameter estimates that would come from different samples drawn from the population that is implicit in the imputation of the missing data.

4.2.7.1 Full Bayesian imputation

Rubin (1987) develops a full Bayesian statistical model for making proper imputations; Schafer (1997a) provides a general approach to the computation of imputed values from this model. If there is a consensual gold standard within the statistical profession for the treatment of missing data, then full Bayesian multiple imputation would seem to be that standard.[23]

To apply this technique to the LAMP data, we used Schafer's (1997b) S-Plus function. 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. Although three to five imputations can suffice, the number of required imputations increases as a function of the amount of missing data. With more than 25 percent of the observations missing on household income, we chose to use 10 imputations.

[23] Western (1999) provides a helpful introduction to Bayesian statistics. The journal issue in which Western's article appears is devoted to substantive examples of Bayesian statistics applied to social scientific research.

4.2.7.2 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).[24] 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 *m*th 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.

It is also possible to use ABB in a manner similar to hotdecking. Suppose *M* bootstrap samples have been drawn. Within each sample, let *W* be a (possibly proper) subset of *Z*, and suppose that $\{W\}$ is a multiway cross-classification over the variables in *W*. For multiple imputation hotdecking, ABB requires that the imputation classes be defined by $\{W\} \times Y$. That is, $\{W\}$ must be stratified by *Y*. With imputation classes so defined, and with *M* bootstrap samples, hotdecking becomes an instance of the approximate Bayesian bootstrap.

[24] Schafer and Schenker (2000) propose a technique that is equivalent to what we describe as conditional mean imputation in section 4.2.5, with the addition of a variance correction. We do not consider this technique here, because it is effectively an algebraic generalization of ABB that short-cuts some of the calculations.

Bayesian bootstrapping requires an algorithm to generate the bootstrap samples (these are available either as features or as contributed macros in a number of standard packages). The imputation model is then estimated separately for each sample, and the analyst assembles the results from the M replicate analyses as described in section 4.2.7.[25]

For the LAMP data we used ABB to generate 10 imputed values for each missing datum, again because more than 25 percent of the observations are missing on household income. Our imputation model consisted of an additive logistic regression of dichotomously defined household income on ethnicity, insurance status, general household well-being and mammography compliance status. This model was estimated on each of the 10 samples bootstrapped from the LAMP data.

For each case missing on household income, we compared fitted probabilities from the regression model with a uniformly distributed random number over the interval 0–1. If the random number was smaller than the fitted probability, missing household income was imputed to be 1; otherwise it was imputed to be 0. Appendix 1, section 7 contains our Stata code for the implementation of ABB.

5. Application of Missingness Techniques to the LAMP Data

We next present the results of applying the eight missingness techniques we have described to the LAMP data. Table 2 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 nonignorable; we do not know the true imputation model; nor are we certain that the substantive model is perfectly

[25] For users of Stata this process is made more straightforward by Paul's (1998) macro.

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.

<u>Insert Table 2 Here</u>

For the LAMP data, several conclusions are apparent:

1. How missing data are treated affects substantive conclusions: 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.
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 2 will not support the conclusion that any missingness technique has performed better than any other. To emphasize the indeterminacy of this examination of the LAMP data, conceivably the case deletion results might be preferable to those of the other methods if the missingness mechanism is nonignorable. Indeed, Allison (2001, p. 7) suggests that case deletion may outperform multiple imputation techniques when missingness is nonignorable. In an attempt to resolve questions of this kind, we turn next to simulations based on the LAMP data.

6. Simulations

We report on two simulation studies. Both are based on the LAMP data; in that sense the simulations are realistic. In the first series, we generated simulated samples in order to study the performance of missingness techniques when dichotomized household income is missing. In the second series, we treated the enthusiasm with which the respondent's doctor supported mammography screening as the variable subject to missingness.

The simulations based on missing household income developed as an outgrowth of the substantive research reported by Fox et al. (1998). Our primary motivation was to assess the extent to which the earlier substantive findings depended on the missingness technique employed (ABB was used by Fox et al. 1998). A secondary motivation was to examine the impact of the choice of missingness technique on coefficients of covariates that had no missingness, given an *X* with missingness that is weakly related to *Y.* We turn next to the household income simulations.

6.1 Household income simulations

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. Using the 857 = 1,119 – 262 complete cases we fit a logistic regression with household income as the response, and compliance status; race/ethnicity; insurance status; and general household well-being as covariates. For the 262 cases missing on household income, we imputed using a procedure analogous to the procedure used in ABB (section 4.2.7.2). Thus, we imputed by comparing random draws over the 0–1 interval with the predicted probabilities from the logistic regression. If, for a given case, the random draw was greater than or equal to the predicted probability, income was imputed to be 1; if less, income was imputed to be 0. The originally nonmissing cases, together with the cases for which household income was imputed, constitute the population for the simulation exercise.

We generated 1,000 fully observed bootstrap samples from the population defined above. Because the within-church intraclass correlation in the original data was quite modest, we did not resample within churches. Thus, we treat each bootstrap sample as though it is a simple random sample.

For each of the 1,000 fully observed bootstrap samples, we created four samples with 262 cases of missingness on household income for a random subsample of observations. The four 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 household income; and MNAR with probability of nonresponse moderately related to household income. Appendix III, section 1, 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 seven missingness techniques. For each of the missingness technique by missingness mechanism combinations we estimated the substantive model for mammography compliance using logistic regression. We did not apply full Bayesian multiple imputation in the household income missingness simulations for data management and programming reasons that are now largely historical.[26]

[26] Although it was feasible to use Schafer's S-Plus function in the one-off analysis of the original data, we carried out our simulation studies using Stata. The use of two statistical systems would have complicated data management of the simulations to an unacceptable degree. For the doctor enthusiasm simulations we wrote our own Stata code for full Bayesian multiple imputation. Because it is not easily generalized, we have not included this code in an appendix.

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 household income as well as whether a respondent is Hispanic is used to create missing data.

Figure 1 summarizes results based on the 28,000 (4 × 7 × 1,000) 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.[27] 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 = 1,000) 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_j th 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

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

[27] 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.

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 Figure 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 *p*th covariate, *j*th missingness mechanism and *k*th 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 *p*th coefficient from the logistic regression fit to the *t*th bootstrap sample subjected to the *j*th missingness mechanism, using the *k*th 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 *p*th covariate is

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

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

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

Column three in Figure 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

$$VI_{pjk} = 100\frac{s_{pjk}^2 - s_p^2}{s_p^2}.$$

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 Figure 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:

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

Insert Figure 1 Here

The results for the simulation of missing household income weakly conform, at best, to the performance expectations of the different missingness techniques under different missingness mechanisms summarized in Table 1. With respect to coefficient bias, casewise deletion is the least adequate performer, except in the MCAR case. Standard errors for casewise deletion are only moderately biased, but the loss of sample size drives variance inflation to a degree of inefficiency matched only by weighted casewise deletion.

Coefficient bias for weighted casewise deletion increases with departure from the MCAR mechanism, with modestly biased standard errors but highly inflated variances. Again, the smaller sample size due to deletion of cases affects efficiency.

The three mean imputation techniques perform roughly identically to each other, and all are essentially unaffected by missingness mechanisms. In contrast, the two stochastic imputation mechanisms perform least adequately. Coefficient bias is greatest for the approximate Bayesian bootstrap, especially when missingness is MCAR. The relatively large average bias in this case is due primarily to an unusual degree of bias in the coefficient of the variable for which missing values were imputed—dichotomized household income. The approximate Bayesian bootstrap also has the largest standard error bias, although for both stochastic imputation techniques variance inflation is nil.

In sum, the results of the household income simulations are puzzling, because the techniques we expected to provide the best results performed relatively poorly. We were concerned that these results were just a fluke, or were in part a strange artifact of the real covariance structure of income with the other variables in the analysis. To alleviate both these fears, we decided to increase the number of simulation iterations, and to simulate missingness on doctor's enthusiasm for mammography screening instead of on income. Fox et al. (1994) note that physician enthusiasm for mammography is the single strongest predictor of mammography screening compliance. In our earlier LAMP data analysis (Table 2), this variable has a sizable and significant effect regardless of missingness technique, which is not the case for income.

6.2 Doctor enthusiasm simulations

For the physician enthusiasm missingness simulations we followed the strategy outlined for household income missingness, with these differences: (i) the number of simulations was increased to 5,000; (ii) the MNAR cases were increased from two to three—"low," "medium," and "high;" and (iii) the list of missingness techniques was extended to include full Bayesian imputation. Appendix III, section 2, supplies details specific to the construction of the simulations for doctor enthusiasm. In the balance of this subsection we describe results for different missingness techniques, and briefly discuss those for mean imputation with a dummy. In the following section, we discuss the results for case deletion and Bayesian multiple imputation.

Figure 2 summarizes results based on the 200,000 (5×8×5,000) regressions using the summary statistics previously described for Figure 1. In these physician enthusiasm missingness 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 setup than it is an indication of the inherent properties 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 III, 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.[28] Presumably the result observed here for mean imputation with a dummy will occur in other situations in 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. Both perform well, but with respect to coefficient bias, they do not exceed the performance of conditional mean imputation and hotdecking.

[28] 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.

7. Discussion of doctor enthusiasm simulations

7.1 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. We had the same expectation, rechecked our work, and found no error that could account for the result. 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. From this we infer that other substantive models, promulgated by other researchers for other problems and other data, are also susceptible to misspecification. We conjecture that application of case deletion will not often lead to favorable results.

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 Figure 2. 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. We infer that the model misspecification is not the result of omitted interactions, but omitted variables. In the realm of nonexperimental research it is difficult to think of a more common misspecification problem than the omission of relevant covariates.

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 Figure 2 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.

7.2 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 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 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 (1997) 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 over-estimation 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.

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. For example, had we been imputing from a fully interacted log-linear model based on $X \mid \{Y \times Z^*\}$, 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.*

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, 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?”

8. Conclusion

Based on our simulation analysis we find that casewise deletion is particularly vulnerable to imperfections in the substantive model. Even in cases where missingness is thought to be MAR and the assumptions of case deletion appear to be satisfied, imperfections in the substantive model that are commonly viewed as minor by substantive researchers can result in considerable coefficient bias. Weighted casewise deletion in our simulations performs well with respect to coefficient bias in the MAR case, but performance deteriorates as missingness becomes MNAR. As with casewise deletion, the simulations demonstrate the inefficiency of weighted casewise deletion. Unless there is strong reason to believe that missingness is MCAR in a given situation, the simulations suggest that neither form of casewise deletion should be a first choice.

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. 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.

This problem of possible imputation model misspecification is not easily solved. A technique which purports to take care of imputation modeling for the analyst by making use of all available covariates and their relationships (Bayesian multiple imputation) can exceed the limits of the data by creating a situation with extensive semi-complete separability. If the model is constrained so that it does not over-tax the data, there is then a risk that the imputation model is incorrect.

What, then, should we do about missing data? Our simulation results suggest that there are important issues to be resolved in the implementation of the leading technical solution via imputation. These results also suggest that conditional mean imputation is not easily dismissed. That we have examined only an extremely limited subset of the potential situations posed by real data cannot, in our view, be taken as grounds for ignoring our findings.

References

Allison, Paul D. 2001. *Missing Data.* Thousand Oaks, CA: Sage Publications.

Anderson, A.B., Basilevsky, A., & Hum, D.P.J. 1983. "Missing Data: A Review of the Literature." in Rossi, Wright & Anderson (eds.) *Handbook of Survey Research*. New York: Academic Press.

Breen, N. and L. Kessler. 1994. "Changes in the use of screening mammography: Evidence from the 1987 and 1990 National Health Interview Surveys." American Journal of Public Health, 84:62-72.

Brick, J.Michael. and Graham Kalton. 1996. "Handling missing data in survey research." *Statistical Methods in Medical Research*, vol. 5, pp. 215-238.

Cochran, William G. 1977. *Sampling Techniques.* New York: John Wiley.

Cohen, Steven B. 1997. "An Evaluation of Alternative PC-Based Software Packages Developed for the Analysis of Complex Survey Data." *The American Statistician*, vol. 51, pp. 285-292.

David, Martin, Roderick J.A. Little, Michael E. Samuhel, and Robert K. Triest. 1986. "Alternative Methods for CPS Income Imputation." *Journal of the American Statistical Association,* vol. 81, pp.499-506.

Efron, Bradley and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall: New York.

Fox, John. 1997. *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks, CA: Sage Publications.

Fox, Sarah A., Albert L. Siu, and Judith A. Stein. 1994. "The Importance of Physician Communication on Breast-Cancer Screening of Older Women." *Archives of Internal Medicine,* 1994, vol. 154, pp. 2058-2068.

Fox, Sarah A., Kathryn Pitkin, Christopher Paul, Sally Carson, and Naihua Duan. 1998. "Breast Cancer Screening Adherence: Does Church Attendance Matter?" *Health Education & Behavior,* vol. 25, pp. 742-758.

Greene, William. 2000. *Econometric Analysis,* Fourth Edition. Upper Saddle River, NJ: Prentice Hall.

Groves, Robert M., Eleanor Singer, and Amy Corning. 2000. "Leverage-Saliency Theory of Survey Participation." *Public Opinion Quarterly*, vol. 64, pp. 299-308.

Heckman, James. 1976. "The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a simple estimator for such models." *The Annals of Economic and Social Measurement*, vol. 5, pp. 475-492.

Heckman, James. 1979. "Sample selection bias as a specification error." *Econometrica*, vol. 47, pp. 153-161.

Horton, Nicholas J., and Stuart R. Lipsitz. 2001. "Multiple imputation in practice: comparison of software packages for regression models with missing variables." *The American Statistician*, vol. 55, pp. 244-254.

Huber, Peter .J. 1967. "The behavior of maximum likelihood estimates under non-standard conditions." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 211-233.

Jones, Michael P. 1996. "Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression." *Journal of the American Statistical Association*, vol. 91, pp.222-230.

Little, Roderick J.A. 1992. "Regression with missing X's: A Review." *Journal of the American Statistical Association*, vol. 87, pp.1227-1238.

Little, Roderick J.A. and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data, 2nd edition*. John Wiley & Sons: New York.

McCullagh, Peter, and John A. Nelder. 1989. *Generalized Linear Models*, Second Edition. New York: Chapman and Hall.

Murray, David M. 1998. *Design and Analysis of Group-Randomized Trials*. New York: Oxford University Press.

Nordholt, Eric S. 1998. "Imputation: methods, simulation experiments and practical examples." *International Statistical Review*, vol. 66, pp. 157-180.

Paul, Christopher. 1998. *Logistic Regression for Data Including Multiple Imputations*. *Stata Technical Bulletin* (STB-45), September, pp. 28-30.

Rao, J.N.K. and Jun Shao. 1992. "Jackknife variance estimation with survey data under hot deck imputation." *Biometrika*, vol. 79, pp.811-22.

Robins, James M., Andrea Rotnitzky and Lue Ping Zhao. 1994. "Estimation of regression coefficients when some regressors are not always observed." *Journal of the American Statistical Association*, vol. 89, pp. 846-866.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons: New York.

Rubin, Donald B. 1996. “Multiple Imputation after 18+ Years.” *Journal of the American Statistical Association*, vol. 91, pp.473-489.

Rubin, Donald B. and Nathaniel Schenker. 1986. “Multiple imputation for interval estimation from simple random samples with ignorable nonresponse.” *Journal of the American Statistical Association*, vol. 81, pp. 366-374.

Rubin, Donald B. and Nathaniel Schenker. 1991. “Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine*, vol. 10, pp. 585-598.

Schafer, Joseph L. 1997a. *Analysis of Incomplete Multivariate Data.* Chapman & Hall: London.

Schafer, Joseph L. 1997b. Software for Multiple Imputation. [http://www.stat.psu.edu/~jls/misoftwa.html].

Schafer, Joseph L. 1999. “Multiple imputation: a primer.” *Statistical Methods in Medical Research*, vol. 8, pp.3-15.

Schafer, Joseph L and Nathaniel Schenker. 2000. “Inference with Imputed Conditional Means.” *Journal of the American Statistical Association*, vol. 95, pp. 144-154

Tanner, Martin A. and Wing Hung Wong. 1987. “The Calculation of Posterior Distributions by Data Augmentation (with discussion).” *Journal of the American Statistical Association*, vol. 82, pp. 528-550.

Western, Bruce. 1999. “Bayesian Analysis for Sociologists.” *Sociological Methods & Research*, vol. 28, pp. 7-34.

Zambrana, Ruth, Nancy Breen, Sarah A. Fox, and Mary Lou Gutierrez-Mohamed. 1999. "Use of Cancer Screening Practices by Hispanic Women: Analyses by Subgroup." *Preventative Medicine*, 29:466-477

Table 1. Apparent Properties of Techniques for Missingness by Mechanisms of Missingness

		Mechanism*		
Technique		MCAR		MAR
Casewise deletion	*b*:	Unbiased	*b*:	Unbiased under certain conditions
	SE(b):	Valid for sample of reduced size	*SE(b)*:	Valid for sample of reduced size
	Var(b):	Inflated by sample reduction	*Var(b)*:	Inflated by sample reduction
Weighted casewise deletion	*b*:	Unbiased	*b*:	Unbiased
	SE(b):	Valid for sample of reduced size	*SE(b)*:	Valid for sample of reduced size
	Var(b):	Inflated by sample reduction	*Var(b)*:	Inflated by sample reduction; uneven weights can also increase variance
Mean imputation	*b*:	Biased	*b*:	Biased
	SE(b):	Biased downward	*SE(b)*:	Biased downward
	Var(b):	Deflated	*Var(b)*:	Deflated
Mean imputation with dummy	*b*:	Potentially biased, extra coefficient	*b*:	Potentially biased, extra coefficient
	SE(b):	Biased downward	*SE(b)*:	Biased downward
	Var(b):	Deflated	*Var(b)*:	Deflated
Conditional mean imputation	*b*:	Nearly unbiased	*b*:	Nearly unbiased
	SE(b):	Biased downward	*SE(b)*:	Biased downward
	Var(b):	Deflated	*Var(b)*:	Deflated
Hotdeck	*b*:	Unbiased	*b*:	Unbiased
	SE(b):	Biased downward	*SE(b)*:	Biased downward
	Var(b):	Inflated	*Var(b)*:	Inflated
Approximate Bayesian bootstrap	*b*:	Unbiased	*b*:	Unbiased
	SE(b):	Unbiased	*SE(b)*:	Unbiased
	Var(b):	Accurate	*Var(b)*:	Accurate
Full Bayesian	*b*:	Unbiased	*b*:	Unbiased
	SE(b):	Accurate	*SE(b)*:	Unbiased
	Var(b):	Accurate	*Var(b)*:	Accurate

*Nonignorable missingness mechanism: All techniques produce biased coefficients when missingness is nonignorable.

Table 2. Logistic Regression of Mammography Compliance Status Under Various Treatments of Missing Data for Household Income[a]

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Mean	Conditional			
	Case	Weighted	Mean	Imputation	Mean		Bayesian	
	Deletion	Case Deletion	Imputation	with Dummy	Imputation	Hot Deck	Bootstrap	Bayesian
Constant	-.49	-.50	-.38	-.38	-.49	-.49	-.38	-.38
	(-1.92)	(-1.94)	(-1.49)	(-1.48)	(-1.99)	(-2.07)	(-1.55)	(-1.52)
MD Enthusiasm	.77	.71	.89	.90	.89	.88	.89	.88
	(4.16)	(3.23)	(6.17)	(6.22)	(6.08)	(5.97)	(6.12)	(6.05)
HH Income > $10,000	.26	.39	.15	.14	.29	.29	.26	.35
	(1.22)	(1.87)	(.68)	(.67)	(1.36)	(1.55)	(1.27)	(1.81)
High School Graduate	.31	.32	.53	.54	.48	.50	.51	.49
	(1.25)	(1.18)	(2.39)	(2.43)	(2.12)	(2.25)	(2.25)	(2.14)
MD Asian[b]	-.26	-.31	-.37	-.37	-.38	-.37	-.38	-.38
	(-1.32)	(-1.69)	(-2.17)	(-2.16)	(-2.20)	(-2.17)	(-2.18)	(-2.18)
MD Hispanic[b]	-.22	-.28	-.56	-.57	-.56	-.57	-.57	-.57
	(-.71)	(-.81)	(-2.56)	(-2.57)	(-2.53)	(-2.53)	(-2.55)	(-2.52)
Same MD 1+ years	.58	.49	.49	.49	.49	.49	.49	.49
	(3.14)	(2.53)	(2.66)	(2.66)	(2.62)	(2.63)	(2.63)	(2.62)
Married	.22	.25	.30	.30	.28	.29	.29	.27
	(1.60)	(1.73)	(2.54)	(2.54)	(2.39)	(2.46)	(2.39)	(2.29)
Medical Insurance	1.19	.77	.90	.90	.88	.88	.88	.88
	(3.88)	(2.57)	(2.95)	(2.95)	(2.91)	(2.88)	(2.88)	(2.88)
Hispanic	-.27	-.32	-.45	-.47	-.42	-.42	-.42	-.41
	(-.94)	(-1.11)	(-1.69)	(-1.71)	(-1.54)	(-1.53)	(-1.55)	(-1.49)
Missingness Dummy				.05				
				(.36)				
N	857	857	1,119	1,119	1,119	1,119	1,119	1,119

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

Note: 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.

[a] The response variable is defined as $Y = 1$ if the respondent is in compliance, = 0 otherwise.

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

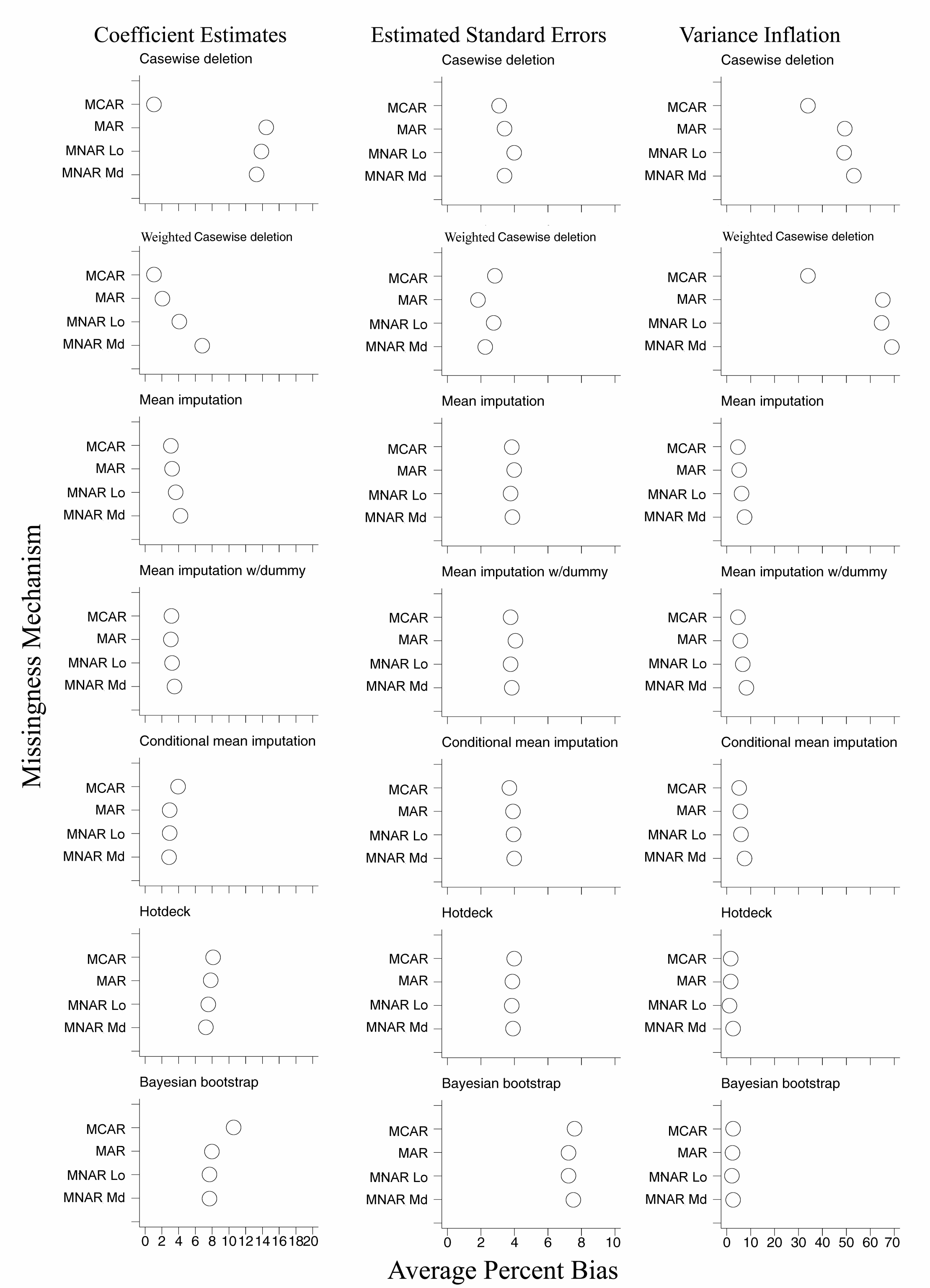

Figure 1. Observed Bias of Estimates from Simulation of Missing Income (*N*=1000)

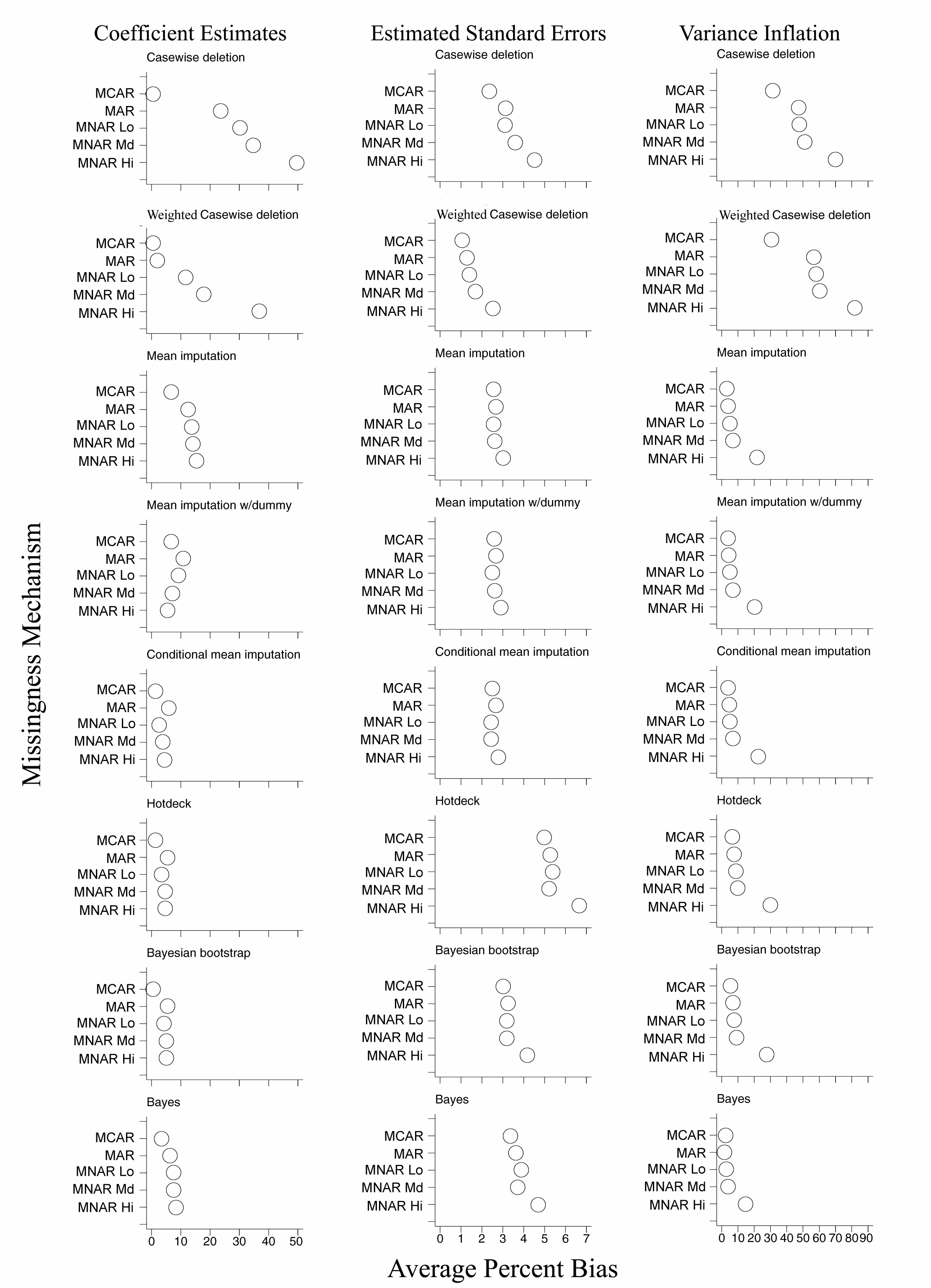

Figure 2. Observed Bias of Estimates from Simulation of Missing MD Enthusiasm (*N*=5000)

Appendix I: STATA code for various technique implementations

All programs assume that the data are loaded, and the following options are in place:

```
#delimit ;
version 5.0;
```

For reference, variables in use throughout Appendix I:
newrace = a race/ethnicity variable, coded white, black, or Hispanic.
goodhse = indicator whether or not respondent reported “good” or better on a question about the general financial well-being of their household without actual dollar amounts.
noinsur = indicator that respondent had no insurance.
comply = indicator of mammography compliance status (1=yes, 0=no).
hispvb_w = indicator, respondent is Hispanic.
partner = indicator, respondent lives with partner or spouse.
dr1_more = indicator, respondent has seen same physician for one year or more.
dr_hisp = respondent's physician is Hispanic.
dr_asian = respondent's physician is Asian.
dropout = respondent is a high school dropout.
inc10 = respondent's household income is > $10,000.
enthu = respondent's physician is enthusiastic about mammography.
chid = the church id number assigned to each church; data were sampled within church clusters.

2. Weighted Casewise Deletion

```
tempfile miss2 miss2x ;

sort newrace goodhse noinsur;
save `miss2', replace ;

g obser = 1 if inc10 ~=.;
g miss = 1 if inc10==.;

collapse (sum) obser miss , by (newrace goodhse noinsur);

g cell = 1 + (miss/obser);
drop obser miss;
sort newrace goodhse noinsur;
save `miss2x', replace;
use `miss2', clear;
merge newrace goodhse noinsur using `miss2x';

logit comply hispvb_w noinsur partner dr1_more dr_hisp dr_asian
dropout inc10 enthu [pweight=cell], cluster(chid)   ;

end;
```

3. Mean Imputation

```
sum inc10 ; replace inc10 = _result(3) if inc10==.;

logit comply hispvb_w noinsr partner dr1_more dr_hisp dr_asian
dropout inc10 enthu ,cluster(chid)   ;

end;
```

4. Mean Imputation with a Dummy

```
* the dummy will be called “misser”;
g misser = 0;
replace misser = 1 if inc10==.;

sum inc10 ; replace inc10 = _result(3) if inc10==.;

logit comply hispvb_w noinsr partner dr1_more dr_hisp dr_asian
dropout inc10 enthu misser ,cluster(chid) ;

*leaves an extra variable in the regression;

end;
```

5. Conditional Mean Imputation (Regression)

```
xi: logit inc10  i.newrace dropout noinsur goodhse  ,cluster(chid)  ;

predict predent;

replace inc10=predent if inc10==. ;

logit comply hispvb_w noinsr partner dr1_more dr_hisp dr_asian
dropout inc10 enthu ,cluster(chid)     ;

end;
```

6. Hotdeck

```
* Because we are hotdecking a binary (0,1) variable, we needn't actually draw within strata, just
calculate the relative frequencies of 0s and 1s in each strata, then randomly assign values based
on those probabilities (mathematically equivalent to actual hot-deck drawing);

tempfile miss2 miss2x ;

sort newrace goodhse noinsur;
save `miss2', replace ;

g obser = 1 if inc10 ~=.;
g inctemp = 1 if inc10==1;

collapse (sum) obser inctemp , by (newrace goodhse noinsur);

g prob = (inctemp/obser);
drop obser inctemp;
sort newrace goodhse noinsur;
save `miss2x', replace;
use `miss2', clear;
merge newrace goodhse noinsur using `miss2x';

qui g shooter = uniform();
replace inc10 = 1 if shooter <=prob & inc10==.;
replace inc10 = 0 if shooter > prob & inc10==.;

logit comply hispvb_w noinsr partner dr1_more dr_hisp dr_asian
dropout inc10 enthu ,cluster(chid);

end;
```

7. Approximate Bayesian Bootstrap

```
*First  we write a little program that will iterate the imputations.  It needs to be in memory before
the data are called and the rest of the code is run;

qui program define dogo ;
version 5.0;
local I = 1 ;
while `I' < 10  {;
use bayes, clear;

bsample ;

xi: logit inc10 i.newrace goodhse noinsur comply, cluster(chid) ;
use bayes, clear;
predict inc10`I'y;
replace shooter = uniform();
replace inc10`I'y = 0 if inc10`I'y <= shooter;
replace inc10`I'y = 1 if inc10`I'y >  shooter;
replace inc10_0`I' = inc10`I'y if inc10==.;
save bayes, replace;
local I = `I' + 1   };
end;


*Now that that is ready, we can begin;
*add appropriate data call here;
use LAMP.dta, clear;

qui g shooter = uniform();
xi: logit inc10 i.newrace goodhse noinsur comply, cluster(chid) ;
* this initial line primes all the variables for the little program above, which is called in two more
lines;
save bayes, replace  ;

qui dogo;

* We would be done imputing at this point, but the 10th iteration requires a special syntax for
implogit.  If you do fewer than 10 imputations, this wouldn't be necessary.;

use bayes, clear;
bsample;
xi: logit inc10 i.newrace goodhse noinsur comply, cluster(chid) ;

use bayes, clear;

predict inc10tp;
qui replace shooter = uniform();
replace inc10tp = 0 if inc10tp <= shooter;
replace inc10tp = 1 if inc10tp >  shooter;
replace inc10_10 = inc10tp if inc10==.;

*Now each observation has 10 income variables, inc10_01 through inc10_10.  For fully
observed cases, all 10 are identical.  For cases where income was missing, they (can) vary across
the 10.  Since the data are now imputed, all that remains is to run 10 regressions, capture and
average the coefs, and capture and adjust the errors.  implogit does this for us: ;

implogit comply hispvb_w noinsr partner dr1_more dr_hisp dr_asian
dropout enthu inc10_01      , impno(10) cluster(chid)   ;
```

```
*implogit is an .ado file that can be found in Stata Technical Bulletin #45, and which can be
downloaded from http://www.stata.com ;

end;
```

Appendix II: Consistency of estimates for conditional mean imputation in logistic regression

Consider the special case of a dichotomous outcome Y and two bivariate normal predictors, X_1 and X_2. Assume the logistic regression model, so that

$$\log\left(\frac{E(y \mid X_1, X_2)}{1 - E(y \mid X_1, X_2)}\right) = b_0 + b_1 X_1 + b_2 X_2,$$ and the

$$\begin{aligned}
E(y \mid X_1) &= \int \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2)}} dF_{X_2|X_1} \\
&= \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2^* E(X_2|X_1))}} \\
|b_2^*| &< |b_2|
\end{aligned}$$

Thus, fitting a logistic regression model to X_1 and $E(X_2|X_1)$ will yield an inconsistent estimate of b_2.

Appendix III: Simulation Details

1. Simulations of missingness on income

Step 1: We used the following procedure to generate data that are similar to the LAMP data:

We began with the 1119 observations of LAMP data that are complete except for income. using the 857 (1119-262) complete cases we fit a logistic regression model predicting income using compliance status, ethnicity, insurance status, and general household financial well-being. This produced a fitted value (fitted probability) prediction ranging from 0-1 for each observation. We filled in the missing income data in the LAMP data by drawing a random 0-1 variable and then comparing it with the predicted probability from the logistic regression model; if the random draw was greater than or equal to the predicted probability, income was imputed to be "1", if less, "0". This procedure was analogous to the procedure used to impute values using the approximate Bayesian bootstrap draw (see Section 4.2.7.2 in the main text). For purposes of this simulation, we treated this completed LAMP data as "the population".

Step 2: We generated 1000 fully observed samples by drawing a random sample of 1119 observations with replacement (bootstrap sample; see for example Efron & Tibshirani, 1993) from this completed LAMP "population" data (created in step 1). Although the data were originally sampled within churches, we did not resample within churches. Rather we ignored churches when drawing our bootstrap samples. Although this eliminates some of the structure of our LAMP data, it simplifies the calculations and the interpretation of results by making the data behave as if they were a simple random sample.

Step 3: For each of the 2000 fully observed samples, we created four samples with missing data by setting income to missing for a random subsample of observations. For each sample we set income to missing for exactly 262 observations (the number of cases missing income in the original LAMP data). The four samples correspond to four different missing data mechanisms: a) missing completely at random; b) missing at random; c) nonignorable missing data, probability of nonresponse is weakly related to income (nonignorable low); and d) nonignorable missing data, probability of nonresponse is moderately related to income (nonignorable moderate).

Generating missing completely at random samples

For our MCAR samples, we chose a simple random sample of 262 observations and set income equal to missing for each observation.

Generating missing at random samples

For our MAR samples, we allowed the probability of missing income to differ by ethnic group. We did this by drawing a stratified random sample, where our strata were African Americans, Hispanics and non-Hispanic whites. Within each strata the sampling rate was held constant at the proportion of nonresponders in the original LAMP data. For example, 56 percent of Hispanics did not report income in the original LAMP data and so for each simulated sample, we chose 56 percent of the Hispanic observations to receive missing income. If the simulated sample had 223 Hispanic observations, then 123 of them received missing income.

Generating nonignorable missing data

To generate nonignorable missing data we allowed the probability of missing income to differ by ethnic group and income. For each ethnic group we split the data into high or low income according to the observed values of the income variable. We then oversampled the observations in the low income strata to receive missing income. To create a sample where nonresponse was weakly related to income (nonignorable low), we set the odds of selection for receiving missing income to be 1.5 times greater for low-income respondents than for high-income respondents. For example, in the completed "population" data set, there were 194 (17.3%) respondents with low income, and 925 (87.7%) with high income. If missingness did not depend on income, we would have chosen approximately 45 respondents with low income (17% of 262) and 217 respondents with high income (83% of 262) to have income set to missing. Instead, we set 60 respondents' with low income (23% of 262) and 202 respondents' with high income (77% of 262) incomes equal to missing to achieve the desired odds ratio of 1.5. To create a sample where nonresponse was moderately related to income, we set the odds of selection for receiving missing income to be 2.0 times greater for low-income respondents than for high-income respondents.

Step 4: For each of the 4000 simulation data sets (1000 simulation iterations X 4 mechanisms of missingness) we applied seven techniques for dealing with missing data (all of those introduced in section 4.2 with the exception of full Bayesian multiple imputation) on each of those samples. Techniques that call for an imputation model used the models described in the sub-sections of section 4.2.

2. Simulation of missing physician enthusiasm for mammography

Our simulations of missing on physician enthusiasm for mammography follow the same general form as the simulations for missing income (described in this appendix, section 1) with the following differences or notes:

Step 1: The data were completed to make the "population" in exactly the same way.

Step 2: Fully observed samples were created in the same way. However, to make the physician enthusiasm simulation more robust, we drew 5000 fully observed samples instead of 1000.

Step 3: In the second set of simulations physician enthusiasm for mammography was made missing instead of income. Since missingness on physician enthusiasm was considerably less common than on income in the actual LAMP data, we generated missing enthusiasms at the same rate as the missing incomes; i.e. 262 missing values per data set of 1119. The MCAR samples were generated as above in this appendix (section 1, step 3), by randomly choosing 262 respondents to have their physician enthusiasm status “lost.” Likewise, MAR samples were drawn to depend on ethnicity, with the same ratios as above (appendix III, section 1, step 3). Non-ignorable samples were generated as above, with missingness dependant on physician enthusiasm and ethnicity. To observe the effects of more extreme levels of non-ignorability, three different levels of nonignorability were imposed. Odds ratio of selection for missingness 1.5 (low nonignorability), Odds ratio 2.0 (moderate non-ignorability), and 4.0 (high non-ignorability), which in practice meant that the odds of selection for missingness were almost perfectly determined by physician enthusiasm and ethnicity (it is almost as if we sampled cases to receive missing income only from the low enthusiasm strata).

Step 4: For each of the 25,000 data sets (5000 iterations X 5 mechanisms of missingness), we applied 8 techniques, as discussed above in section 4.2. Section 4.2 discusses the application of each technique for data that are missing income. To apply to missing enthusiasm, all imputations and replacements were changed to pertain to enthusiasm, and, where applicable, used models more appropriate to predicting physician enthusiasm than income. The models corresponding to each technique, and the section number of the presentation of the technique for the missing income case (from the main text above) are listed below:

Weighted casewise deletion (casewise re-weighted)—4.2.2

In the missing enthusiasm simulations, 6 imputation classes were used, based on ethnicity (newrace, 3 categories) and high school completion (dropout, 2 categories).

Mean imputation—4.2.3 & Mean imputation with a dummy—4.2.4

In the missing enthusiasm simulations, the mean for physician enthusiasm was imputed for all missing values.

Conditional mean imputation—4.2.5

In the missing enthusiasm simulations, conditional means were calculated based on 6 imputation classes derived from ethnicity (newrace, 3 categories) and high school completion (dropout, 2 categories).

Hotdeck imputation—4.2.6

In the missing enthusiasm simulations, hotdeck draws were taken within 6 imputation classes based on ethnicity (newrace, 3 categories) and high school completion (dropout, 2 categories).

Full Bayesian multiple imputation—4.2.7.1

We wrote our own Stata routine to implement Schafer's (1997b) algorithm for full Bayesian imputation in the simulations of missing enthusiasm. Schafer's model calls for the inclusion of all model covariates, including the outcome, so we included all variables in our imputation model.

Approximate Bayesian bootstrap—4.2.7.2

To apply the ABB to the simulations of missing physician enthusiasm, we used all other model covariates (including the outcome) to predict physician enthusiasm for mammography.