Annu. Rev. Public Health 2004. 25:99–117
doi: 10.1146/annurev.publhealth.25.102802.124410

WHAT DO WE DO WITH MISSING DATA? SOME OPTIONS FOR ANALYSIS OF INCOMPLETE DATA

Trivellore E. Raghunathan
Department of Biostatistics and Institute for Social Research,
University of Michigan, Ann Arbor, Michigan 48109; email: teraghu@umich.edu

■ **Abstract** Missing data are a pervasive problem in many public health investigations. The standard approach is to restrict the analysis to subjects with complete data on the variables involved in the analysis. Estimates from such analysis can be biased, especially if the subjects who are included in the analysis are systematically different from those who were excluded in terms of one or more key variables. Severity of bias in the estimates is illustrated through a simulation study in a logistic regression setting. This article reviews three approaches for analyzing incomplete data. The first approach involves weighting subjects who are included in the analysis to compensate for those who were excluded because of missing values. The second approach is based on multiple imputation where missing values are replaced by two or more plausible values. The final approach is based on constructing the likelihood based on the incomplete observed data. The same logistic regression example is used to illustrate the basic concepts and methodology. Some software packages for analyzing incomplete data are described.

Key Words available-case analysis, observed data likelihood, missing data mechanism, multiple imputation, nonresponse bias, weighting

INTRODUCTION

Missing data are a ubiquitous problem in public health investigations involving human populations. In a cross-sectional study relying on a survey, subjects may refuse to participate entirely or may not answer all the questions in the questionnaire. The former type of missing data is called unit nonresponse, and the latter, item nonresponse. In a longitudinal study, subjects may drop out, be unable, or refuse to participate in subsequent waves of data collection. The missing data in this context may be viewed as unit or item nonresponse. For instance, in a cross-sectional analysis of data from a particular wave, drop-outs may be viewed as unit nonrespondents, whereas in a longitudinal analysis involving data from all waves, missing data due to drop-outs may be viewed as item nonresponse.

The standard approach, a default option in many statistical packages, is to restrict the analysis to subjects with no missing values in the specific set of variables. This so-called available-case analysis can yield biased estimates. Sometimes, where multiple analyses are involved, including descriptive analysis, the standard approach excludes subjects with any missing values in any of the variables used in at least one analysis (the so-called complete-case or listwise deletion analysis). Other ad hoc approaches such as treating the missing data as a separate category also result in biased estimates (35).

To demonstrate the potential bias in the available-case estimates, consider a population with a binary disease variable, D, a binary exposure variable, E, and a continuous confounder, x. Suppose that this population adheres to the following model assumptions:

$$x \sim N(0, 1),$$

$$\mathit{logit}\,\Pr(E = 1|x) = 0.25 + 0.75 \times x, \text{ and}$$

$$\mathit{logit}\,\Pr(D = 1|x, E) = -0.5 + 0.5 \times E + 0.5 \times x.$$

This model and the parameters were partly motivated by the analysis of cohort data. A sample size of 1000 is drawn from essentially an infinite population, and a logistic regression model is fitted, with D as the outcome variable and x and E as predictors (the last equation in the above model assumptions), resulting in the estimates of -0.54, 0.57, and 0.53 for the intercept, regression coefficients for E and x, respectively. These estimates are close to the true values in the logistic regression model given above.

Now suppose that some values of x are deliberately set to be missing using the following logistic model,

$$\mathit{logit}[\Pr(x \text{ } \mathit{is\ missing})] = -1.11 - 1.09 \times D - 1.85 \times E + 2.31 \times D \times E.$$

That is, for each subject, generate a uniform random number between 0 and 1, and if this number is less than or equal to the probability computed from the above equation, then set that value of x to missing. This model for deleting values was designed to produce approximately 15% missing values, and the percentage of observations with missing values in the 4 cells formed by D and E was 26% for $(D = 0, E = 0)$, 11% for $(D = 1, E = 0)$, 5% for $(D = 0, E = 1)$, and 14% for $(D = 0, E = 0)$. These parameters were chosen to match the missing values in the cohort data that motivated this experiment. The same logistic regression model was fitted to the data set, now restricted only to those subjects with no missing values in x. The resulting estimates of intercept, regression coefficients for E and x were -0.30, 0.28, and 0.52, respectively. The estimate of the regression coefficient for E is remarkably different when compared to the "before-deletion" estimate as well as the true value of 0.5. One might wonder whether the observed difference is due to idiosyncrasies of this particular data set or has occurred purely by chance.

To investigate further, the experiment was replicated as follows:

1. Generate 2500 samples (we will call these before-deletion samples) each of size 1000.

2. Delete some values of x in each data set using the same logistic model mechanism given above (we will call the corresponding data sets with values of x set to missing as "after-deletion" data sets).
3. Fit logistic regression models to both 2500 before-deletion and the corresponding after-deletion data sets.
4. The primary parameter of interest is the regression coefficient for E, the log-odds ratio measuring the association between D and E adjusted for x. Figure 1 provides the histogram of 2500 estimated regression coefficients from before-deletion data sets, and Figure 2 provides the histogram of the corresponding 2500 available-case estimates from after-deletion data sets. Figure 1 is centered on the true value 0.5 and is normal in shape, given the large sample size ($n = 1000$). Figure 2, on the other hand, is centered approximately at 0.20, and in fact, the true value lies in the tail of the sampling distribution. This basic simulation study demonstrates that the standard practice of omitting subjects with any missing values can be invalid.

This article reviews three approaches for correctly analyzing incomplete data, and these will be evaluated using the same simulated data sets used to illustrate the

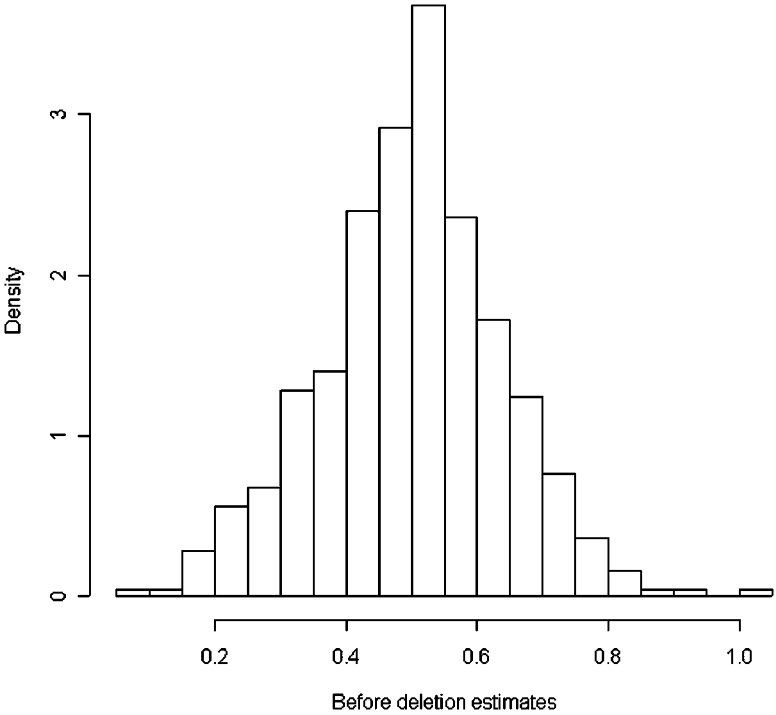

Figure 1 Histogram of logistic regression coefficient from 2500 simulated data sets before deleting any covariate values.

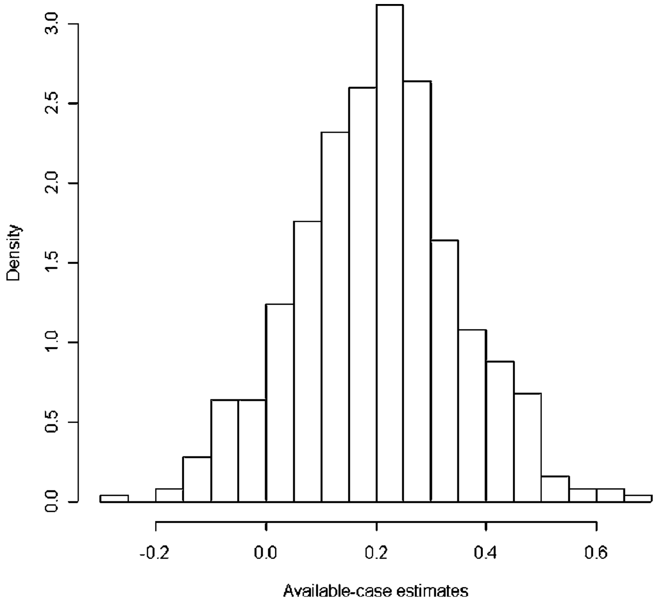

Figure 2 Histogram of logistic regression coefficient from 2500 simulated data sets after deleting some covariate values and using available-cases.

perils of using the available-case approach. The first approach involves attaching weights to each subject included in the analysis to represent subjects who were excluded. This is often used to compensate for unit nonresponse in surveys [see (9) for a review]. The second approach is through multiple imputation (29, 30), where the missing set of values is replaced by more than one plausible set of values. Each plausible set of values in conjunction with the observed data results in a completed data set. Each completed data set is analyzed separately using the standard complete data software, and the resulting point estimates and standard errors are combined using a simple formula described later.

The distinction between the observed and filled-in values must be incorporated in any subsequent analysis of data with imputed values. That is, the filled-in values for any one subject with missing values should not be considered as microdata for that subject but rather as values that are statistically plausible given other information on that subject. These filled-in or completed data sets are plausible samples from the population under certain assumptions. Thus, the completed data sets should result in inferences (point estimates and confidence intervals, for example) that are within the realm of statistical plausibility of inferences that would have been obtained had there been no missing data. In that respect, the multiple imputation approach is a statistical approach of “rectangularizing” the observed data to exploit the available complete data software to obtain valid inferences. The inferential validity based on the multiple imputed data sets is the goal, and any imputation procedure should not viewed as a method for recovering the missing values for any given individual.

The third approach is based on the likelihood constructed from the observed incomplete data. This approach has a long history: The earliest reference seems to be McKendrick (21), who used an algorithm similar to the Expectation-Maximization (EM) algorithm (5) to obtain estimates from a sample with missing values. The EM algorithm is a popular approach for maximizing the observed data likelihood. This paper emphasizes weighting and imputation approaches and briefly discusses the maximum likelihood approach. The maximum likelihood approach is available only for very few selected models such as multivariate normal linear regression, contingency tables, and certain other multivariate analyses. A detailed discussion of the maximum likelihood approach can be found in Reference 19 and a Bayesian approach similar in spirit in Reference 34.

All these methods make certain assumptions about why data are missing. The answer to this question is stated in terms of assumptions about the missing data mechanism. In the next section, this critical concept is described. The missing data mechanisms comprise probabilistic specifications of why data are missing. These mechanisms were developed by Rubin (28) and later extended to longitudinal data by Little (17). All three approaches described above are valid under the so-called missing at random (MAR) mechanism, and, generally, the available-case approach is valid under a stricter mechanism called missing completely at random (MCAR). The third section describes the weighting approach; the fourth section, the multiple imputation approach; and the fifth section describes the likelihood-based methods. The final section concludes with discussion and limitations.

MISSING DATA MECHANISM

To understand the taxonomy of missing data mechanisms, consider a simple case where the analysis of interest is concerned with a set of variables (U, V), the missing values are in a single variable U, and variables in vector V have no missing values. Our simulation study falls into this category. Suppose that R denotes an indicator variable taking the value 1 if U is observed and 0 if U is not observed. Thus, the observed data are R, (U, V), if $R = 1$ and V, if $R = 0$. The missing data mechanism or, equivalently, the response mechanism is the conditional probability distribution of R, $Pr(R|U,V)$, given (U, V). That is, what probabilistic mechanism governs the observation of U, specifically in relation to the variables of interest U and V?

The missing data mechanism is called MCAR, if $Pr(R = 1|U,V) = c$, a constant. That is, the probability that U is observed is independent on the underlying values of U or V. Under this assumption, the available-cases (that is, with $R = 1$) constitute a random subsample of the original sample. In a more general case, where missing values can also occur in *V*, the MCAR assumption implies that the missing values in any variable are independent of the underlying values of *U* or *V*. The subjects included in the available-case analysis, therefore, constitute a random subsample of the original sample. Thus, the analysis that includes only those who have *U* and *V* observed is generally valid under this assumption because the process of excluding the subjects with any missing values does not distort the representativeness of the original sample. This assumption is clearly violated in our simulation study where the percentages with missing values differ across the four cells based on *D* and *E*. The MCAR is a rather strong assumption and is rarely satisfied in practical applications. Sometimes the available-case analysis may be valid under a weaker assumption [see, for example, (7, 12, 15)], but these exceptions are few and idiosyncratic.

A weaker assumption is MAR where, again, resorting to the simple case first, $Pr(R = 1|U,V) = f(V)$, a function that depends on *V* but not on *U*. The deletion mechanism used in the simulation study falls under this category where the missingness in *x* depends on *D* and *E* but not on *x*. In essence, this assumption entails that for two individuals with the same value of *V*, one has *U* observed and the other has *U* missing; the missing *U* is arising from the same distribution (for a given *V*) as the observed one. That is, conditional on *V*, the missing *U* is predictable from the observed distribution of *U*. In a more general situation where the missing values can be in several variables, the exact analytical specification of this assumption is difficult. Loosely speaking, suppose that $d_{i,obs}$ denotes the observed components of complete data, (U_i, V_i), for subject *i*, and $d_{i,miss}$ denotes the missing components. For two individuals, *i* and *j*, when $d_{i,obs} = d_{j,obs}$, the missing components, $d_{i,miss}$ and $d_{j,miss}$, have the same distribution.

Given this conditional nature of the assumption, the stronger the correlates of *U* in *V*, the weaker is the assumption about the missing data mechanism for *U*. For example, if *U* were income, then having a rich set of variables to condition on in *V*, such as age, gender, education, occupation, property values, monthly expenditures, and neighborhood level information, makes the assumption about missing data considerably weaker when compared to MCAR or when the list of variables to be conditioned on is limited to, say, age and gender. The limitation of the MAR assumption is due to lack of appropriate variables that can be conditioned on in the analysis, and empirically it has been shown to be reasonable in practical situations (4, 33). The three approaches discussed in this paper are valid under different versions of this weaker assumption about the missing data mechanism.

Finally, the missing data mechanism is said to be Not-Missing at Random (NMAR), if $Pr(R = 1|U, V) = f(U, V)$, a function that certainly depends on *U* but may also depend on *V*. That is, even after conditioning on *V*, the distribution of *U* for the respondents and nonrespondents are dissimilar. This function, however, is not estimable from the observed data because whenever $R = 0$, *U* is unobserved. Therefore, an explicit form of *f* has to be specified and the data cannot be used to empirically verify the validity of this assumption.

The specification of $Pr(R = 1|U, V)$, in conjunction with the substantive model $Pr(U, V)$, is used to construct inferences about the parameters of interest. This is a selection model method and was first proposed by Heckman (8). The alternative approach is to specify how different the distributions of (U, V) are for the respondents and nonrespondents. That is, specify the population distribution as a mixture of two components, $Pr(U, V|R = 1)$ and $Pr(U, V|R = 0)$, for respondents and nonrespondents, respectively. This mixture is used to construct inferences about the population. Again, there is no data to specify the part of the mixture $Pr(U, V|R = 0)$ because U is unobserved whenever $R = 0$. This approach was first proposed by Rubin (27, 29) and later extended by Little (16). In any event, both these approaches make empirically unverifiable assumptions, and their use is very limited to situations where some prior knowledge may exist to specify the mixture distribution or the selection model. For more details see Chapter 15 in Reference 19.

WEIGHTING

The origins of weighting may be traced to sample survey practice where it is used as a simple device to account for unequal probabilities of selection (11). The survey weight for an individual is the inverse of his/her selection probability. To be concrete, suppose that a national probability survey sampled 1000 subjects with 500 Whites and 500 African Americans. That is, the African Americans were oversampled with respect to their representation in the population. Suppose that one were to ignore this fact and proceeded to compute a simple average of 1000 observations as an estimate of the population mean. If there are large differences between Whites and African Americans in the survey variable of interest, the sample mean will be a distorted representation of the population mean. Down-weighting the observations on African Americans to their representation in the population and up-weighting the observations on Whites to their representation in the population would obtain the accurate picture of the population. Thus, the weighted average (where the weights are inverse of their selection rate) is an unbiased estimate of the population mean whereas the simple mean is not. However, special software is needed to compute standard errors and confidence intervals that use these weights. Currently, popular packages such as SAS, STATA, and SUDAAN have built-in routines to take into account these survey weights.

Weighting to compensate for nonresponse is an extension of the same idea. That is, excluding the subjects because of missing values is a distortion of the representation in the original sample, and weights are attached to subjects included in the analysis to restore the representation. To be concrete, consider the simulation example. Suppose that n_{de} is the number of subjects with $D = d$ and $E = e$ where $d, e = 0{,}1$. Let the number of subjects with observed x in the corresponding cell be r_{de}. To restore the available-case analysis to its original sample representation, we should weight each respondent in (d, e) cell by $w_{de} = n_{de}/r_{de}$, inverse of the response rate in that cell or the inverse of the selection probability into the data

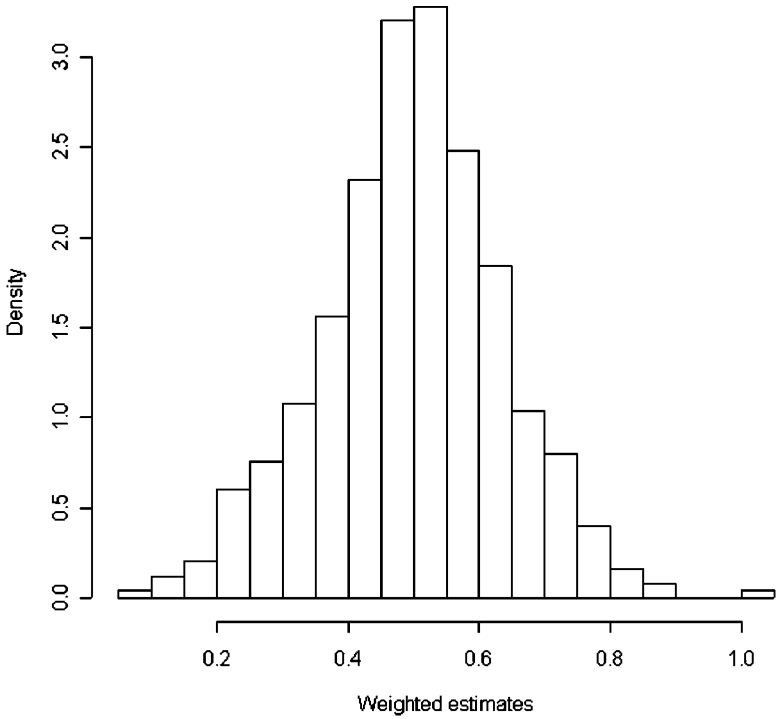

Figure 3 Histogram of logistic regression coefficient from 2500 simulated data sets after deleting some covariate values and using weights to compensate for missing data.

analysis. A weighted logistic regression can be used to estimate the parameter of interest (2).

Figure 3 gives the point estimates for the same 2500 data sets analyzed previously, but this time based on weighted logistic regression where the weights were computed as the inverse of the response rates in the four cells. As can be seen, the weighted estimates are unbiased and, understandably, slightly more variable than the before-deletion estimates.

There are several approaches for constructing weights (9). Consider the example given earlier where the analysis of interest is concerned with a set of variables *(U, V)*, the missing values are in a single variable *U*, and variables in a vector *V* have no missing values. The adjustment-cell method involves constructing a contingency table through cross-classification of sampled subjects based on *V*. The inverse of the response rate in any cell is the weight attached to each respondent in that cell. The underlying assumption is that, conditional on belonging to an adjustment cell, the respondents and nonrespondents have similar distribution of variables with missing values (that is, respondents and nonrespondents are exchangeable within the adjustment cell). The simulation example described above used this method to construct the weights based on D and E. The disadvantage of this approach is that the continuous variables in V have to be categorized, and if V has a large number of variables, the contingency table can be sparse, leading to unstable weights.

An alternative approach is to estimate the response propensity through a logistic regression model, $\Pr(R = 1|V) = [1 + \exp(-\beta_o - V^t\beta_1)]^{-1}$, where β_o and $\beta_1$are the unknown regression coefficients and the superscript t denotes matrix transpose. Suppose that $\hat{\beta}_o$ and $\hat{\beta}_1$ are the estimated regression coefficients; the weights for a respondent j is then defined as $w_j = 1 + \exp(-\hat{\beta}_o - V_j^t\hat{\beta}_1)$. One may include the interaction terms between different variables or transformation in V. There is no need to categorize continuous variables. This response propensity approach is more practical when we have large numbers of variables. Sometimes categorization of estimated response probabilities forms adjustment cells.

MULTIPLE IMPUTATION

Weighting is a simple approach for making the subjects included in the available-case analysis representative of the original sample and is effective in removing nonresponse bias. However, by including only subjects with complete data, it ignores partial information from subjects with incomplete data. For example, in a multiple linear regression with some subjects missing one variable at the most, it can be very inefficient to ignore information on the rest of the variables, especially if they are good predictors of the variables with missing values.

An alternative approach is based on filling in or imputing the missing values in the data set. Again, this approach can be traced back to survey practices adopted by the U.S. Bureau of Census [See (6) for historical accounts]. This practice of filling in the missing values (this is called single imputation method) in the survey practice is attractive for several reasons. First, imputation adjusts for differences between nonrespondents and respondents on variables observed for both and included in the imputation process, as well as differences on variables not included in the model that are predicted by the model; such an adjustment is generally not made by available-case analysis. Second, the complete data software can be used to process the data to obtain descriptive statistics and other statistical measures. This is a significant advantage because complete-data software has kept closer pace with the statistical methodological developments than the incomplete-data software. Third, when a data set is being produced for analysis by the public or multiple researchers, imputation by the data producer allows the incorporation of specialized knowledge about the reasons for missing data in the imputation procedure, including confidential information that cannot be released to the public or other variables in the imputation process that may not be used in substantive analysis by a particular researcher. Raghunathan & Siscovick (24) demonstrate that using an auxiliary variable in the imputation process can improve the efficiency considerably. Moreover, the nonresponse problem is solved in the same way for all users so that analyses will be consistent across users. The researcher using the filled-in data can concentrate on addressing substantive questions of interest and not be distracted by incomplete data. See Reference 31 for a detailed list of applications of this approach.

Although single imputation, that is, imputing one value for each missing datum, enjoys the positive attributes just mentioned, analysis of a singly imputed data set using standard software fails to reflect the uncertainty due to the fact that the imputed values are plausible replacements for the missing values but are not the true values themselves. As a result, such analyses of singly imputed data tend to produce estimated standard errors that are too small, confidence intervals that are too narrow, and significance tests with *p*-values that are too small.

Multiple imputation (29, 30) is a technique that seeks to retain the advantages of single imputation while also allowing the uncertainty due to imputation to be incorporated into the analysis. The idea is to create more than one, say *M*, plausible sets of replacements for the missing values, thereby generating *M* completed data sets. The variation across the *M* completed data sets reflects the uncertainty due to imputation. Typically, *M* is not larger than five.

The analysis of the *M* completed data sets resulting from multiple imputation proceeds as follows:

1. Analyze each completed data set separately using a suitable software package designed for complete data (for example, SAS, SPSS, or STATA).
2. Extract the point estimate and the estimated standard error from each analysis.
3. Combine the point estimates and the estimated standard errors to arrive at a single point estimate, its estimated standard error, and the associated confidence interval or significance test.

Suppose e_l is the estimate and s_l its standard error, based on the completed data set $l = 1, 2, \cdots, M$, where$M \geq 2$. The multiply imputed estimate is the average,

$$\bar{e}_{MI} = \frac{1}{M} \sum_{l=1}^{M} e_l,$$

and the standard error of the multiply imputed estimate is

$$s_{MI} = \sqrt{\bar{u}_M + \frac{M+1}{M} b_M},$$

where

$$\bar{u}_M = \frac{1}{M} \sum_{l=1}^{M} s_l^2 \quad \text{and}$$

$$b_M = \frac{1}{M-1}\sum_{l=1}^{M}\left(e_l - \bar{e}_{MI}\right)^2$$

The sampling variance (term inside the square root sign) has two parts: The first part is the average sampling variance by treating the imputed values as though they are real. This is called within-imputation component of variance. The second part is the variability across the imputed values (the between-imputation component of variance), which is not estimable unless more than one plausible set of values are used as fill-in. Rubin & Schenker (32) and Rubin (30) derived the sampling distribution as t-distribution with degrees of freedom, $\nu = (M-1)(1+r_M)^2$, where $r_M = \bar{u}_M/[(1+M^{-1})b_M]$. Several other methods have been developed to construct intervals (1, 13, 14).

Most straightforward justification for generating more than one plausible set of values is through draws from the predictive distribution of the missing values conditional on the observed values. Revisiting the simulation example, one could generate several plausible values from the predictive distribution based on a regression model,

$$x = \beta_o + \beta_1 D + \beta_2 E + \beta_3 D \times E + \varepsilon,$$

where $\beta = (\beta_o, \beta_1, \beta_2, \beta_3)$ is a vector of regression coefficients, and the residual $\varepsilon \sim N(0, \sigma^2)$. A simple approach is to estimate the regression coefficients and the residual variance using subjects with x, D, and E observed. Suppose $\hat{x}$ is the predicted value for an individual with missing x. Adding different noise variables $z \sim N(0, \hat{\sigma}^2)$, where $\hat{\sigma}^2$ is the estimate of residual variance, to the predicted value generates plausible values. This method is reasonable for large sample size but still is not proper (30) because plausible values are generated without reflecting uncertainty in the estimates of the regression coefficients and residual variances. A proper approach reflects uncertainty in every estimate while generating plausible values. A proper approach for generating the plausible values is the Bayesian approach where the missing values are drawn from the posterior predictive distribution of the missing observations conditional on the observed data. The approach is technical and the procedure for the simulation example is described in the appendix. More formal description of a proper method can be found in References 19, 30, and 34.

The fully Bayesian approach described in the appendix was implemented on 2500 simulated data sets with missing values described earlier. Five imputations were created for each of 2500 data sets with missing values. Five completed data sets were analyzed by fitting a logistic regression model to each. The multiple imputation estimate and its standard error were computed using the formula given above. Figure 4 gives the histogram of 2500 multiple imputation estimates, which shows that the sampling distribution is centered on the true value 0.5.

The most straightforward approach for creating multiple imputations is model-based using a Bayesian formulation. That is, draw values from the posterior

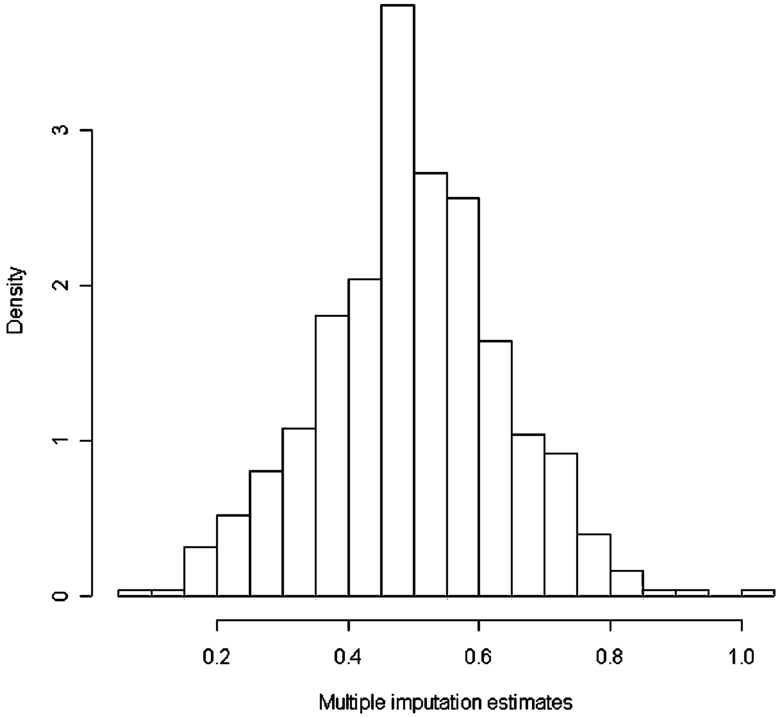

Figure 4 Histogram of logistic regression coefficient from 2500 simulated data sets after deleting some covariate values and using multiple imputation method.

predictive distribution, $\Pr(d_{miss}|d_{obs})$, of the missing values, d_{miss}, conditional on the observed values, d_{obs}. Little & Raghunathan (18) argue that the imputation should condition on as much observed information as possible to make MAR plausible and imputations efficient. Schafer (34) has developed several routines for implementing the Bayesian method where one can achieve approximate normality of continuous variables through transformation and a limited number of categorical variables (see the website http://www.stat.psu.edu/~jls). However, this model-building task can be quite difficult in practical situations with hundreds of variables, skip patterns in the questionnaire, different types of variables such as continuous, categorical, count, and semicontinuous (basically, continuous but with a spike at 0). Other common problems include structural dependencies. For example, the question asking about years of smoking is applicable only to current and former smokers, whereas years since quitting is applicable only to former smokers. Also, years smoked cannot exceed the age of the person.

An alternative, though not a fully model-based Bayesian but fully conditional approach, is the sequential regression approach (23), which builds on a version for continuous variables originally proposed by Kennickel (10). A brief description of SRMI is as follows: Let X denote the fully observed variables, and let $Y_1, Y_2, \ldots, Y_k$ denote k variables with missing values in any order. The imputation process for $Y_1, Y_2, \ldots, Y_k$ proceeds in c rounds. In the first round, Y_1 is regressed on X, and the missing values of Y_1 are imputed (using a process analogous to that described for the logistic regression example in the Appendix); then Y_2 is regressed on X and Y_1 (including the imputed values of Y_1), and the missing values of Y_2 are imputed; then Y_3 is regressed on X, Y_1 and Y_2, and the missing values of Y_3 are imputed; and so on, until Y_k is regressed on $X, Y_1, Y_2 \ldots, Y_{k-1}$, and the missing values of Y_k are imputed.

In Rounds 2 through c, the imputation process carried out in Round 1 is repeated, except that now, in each regression, all variables except for the variable to be imputed are included as predictors. Thus, Y_1 is regressed on $(X, Y_2, Y_3, \ldots, Y_k)$, and the missing values of Y_1 are imputed; then Y_2 is regressed on $(X, Y_1, Y_3, \ldots, Y_k)$, and the missing values of Y_2 are imputed; and so on. After c rounds, the final imputations of the missing values in $(Y_1, Y_2, \ldots, Y_k)$ are used.

An SAS-based software IVEware implementing this approach is available from the website http://www.isr.umich.edu/src/smp/ive. The S-plus version of a similar approach is available from http://www.multiple-imputation.com. For the regressions in the SRMI procedure, IVEware allows the following models:

1. a normal linear regression model if the Y-variable is continuous;
2. a logistic regression model if the Y-variable is binary;
3. a polytomous or generalized logit regression model if the Y-variable is categorical with more than two categories;
4. a Poisson loglinear model if the Y-variable is a count;
5. a two-stage model if the Y-variable is mixed (i.e., semicontinuous). In the first stage zero/nonzero status is imputed using a logistic regression model. Conditional on being nonzero, a normal linear regression model is used to imput a nonzero value.

Because SRMI requires only the specification of individual regression models for each of the Y-variables, it does not necessarily imply a joint model for all of the Y-variables conditional on X. This procedure compares well with the fully Bayesian approach as demonstrated through the simulation study by Raghunathan et al. (23). This is the most practical approach in many situations involving structural dependencies and the large number of predictors of varying types.

MAXIMUM LIKELIHOOD

In the complete data statistical methodology, maximum likelihood for a given model is a dominant inferential procedure, for example, the linear, logistic, Poisson, log-linear, and random effects models. All use likelihood as a basis for constructing inferences. Extending the same notion, one possibility is to base our inferences on likelihood function constructed from the actual observed data set. To motivate this approach, consider an example based on a random sample of size n from a bivariate normal distribution,

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}\right].$$

Suppose that p subjects provide both X and Y, q subjects provide only X and not Y, and r subjects provide Y but not X. The objective is to estimate the five unknown parameters $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY})$. Once these estimates are obtained, any function of these parameters, such as correlation coefficient or regression coefficients, can be computed. Wilks (36) addressed this estimation problem using several approaches including the maximum likelihood method.

The observed data likelihood is a product of three components,

$$L_{obs} = L_p \times L_q \times L_r,$$

where the first component is the contribution from p subjects who provided both X and Y. This is a product of bivariate normal density function evaluated at the observed values for those p subjects, and it involves all five parameters. The second component is the contribution from q subjects who provided only X. This is a product of univariate normal density functions involving only (μ_X, σ_X^2). Finally, the third component, based on r subjects who provided only Y, is a product of univariate normal density functions involving only (μ_Y, σ_Y^2). The observed data likelihood is then maximized with respect to $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \sigma_{XY})$ using some iterative routines such as Newton-Raphson method or, more popular, the EM algorithm (5).

More generally, suppose that Y_i is a complete data vector on subject $i = 1, 2, \ldots, n$; $Y_{i,obs}$ denotes the observed components and $Y_{i,miss}$ the missing components; and$Y_i = (Y_{i,obs}, Y_{i,miss})$. Suppose the complete data model is

$$f(Y_i|\theta) = f(Y_{i,obs}, Y_{i,miss}|\theta),$$

where θ is the unknown parameter to be estimated. The observed data likelihood is

$$L_{obs}(\theta) = \prod_{i=1}^{n} L(\theta|Y_{i,obs}) \propto \prod_{i=1}^{n} \int f(Y_{i,obs}, Y_{i,miss}|\theta)\, dY_{i,miss}.$$

The justification that this is the correct likelihood to be maximized when data are MAR is given in Reference 28.

The approximate sampling variances are typically estimated by inverting the negative second derivative of the logarithm of the observed data likelihood. Though the estimates based on this approach are perhaps the most efficient and enjoy all the nice properties accorded to maximum likelihood estimates, implementation is quite difficult even in the simple logistic regression case considered earlier. Typically special software needs to be developed for a particular problem.

Software for multivariate normal model is available in SPSS 10 and as an add-on package in GAUSS. The general location model, which can be used to model normal continuous and discrete variables simultaneously, was considered by Little & Schluchter (20), Raghunathan & Grizzle (22) and Schafer (34). Software for fitting such models using a Bayesian approach was developed by Schafer (34). Due to these technical difficulties, this method is not practical in many real world applications.

DISCUSSION AND LIMITATIONS

Analysis of data with some missing values is an important problem, and the standard strategy of including only those on whom a particular analysis can be carried out can lead to biased estimates. Three approaches have been discussed with increasing levels of statistical sophistication. Weighting is the simplest approach; multiple imputation is at the second level but is still a more general approach. The maximum likelihood approach is the most difficult, often requiring user-developed software for implementation. All these approaches are valid under a general class of mechanisms called MAR, whereas the available-case analysis is generally valid under MCAR, though there are some exceptions (7, 12). These exceptions are few and idiosyncratic, and are often difficult to verify in a practical setting. Even if the data are MCAR, the available-case method is less efficient owing to discarding subjects with partial information. It is not uncommon for substantial numbers of subjects to be excluded in a regression analysis, even though each subject is missing only a few variables.

Weighting is a simple device to correct for bias, but it suffers from the same disadvantages as the available-case method in terms of efficiency. It still discards partial information from subjects with missing values. Thus, if the bias correction is the motivating factor, then weighting should be used to compensate for missing data. Either the adjustment-cell method or the response propensity method can be used to derive the weights, though the response propensity method relies on the logistic regression model. A compromise might be to use the response propensities to form adjustment cells to rely less on the correctness of the response propensity model specification.

Perhaps the most practical approach is based on multiple imputation. This approach involves an upfront investment in multiply imputing the missing values in the database. Once multiply imputed, any complete data software can be used to repeatedly analyze the completed data sets, extract the point estimates and their standard errors, and combine them using the formula given in the third section of this review. The last step can be carried out using a spreadsheet program such as Excel. There are user-developed routines now available in STATA (3) and IVEWARE (23), and in commercial software such as SAS version 8.2 and SOLAS. Though this method requires additional storage and extra steps of repeated analysis and combining estimates, in the grand scheme of public health investigations, it is a minor step, especially owing to the availability of software for creating multiple imputations.

Though we emphasize multiple imputations, it is possible to correct the standard errors using the single imputation method. For a limited set of variables such as means and proportions and the correlation coefficients, Rao & Shao (26) and Rao (25) have proposed the Jackknife method for computing correct standard errors from singly imputed data sets. Nevertheless, the multiple imputation approach seems to be the most practical approach in a setting involving a large data set with multiple researchers using different portions of the same data set, as well as for a single researcher analyzing a particular data set with missing values, provided an upfront investment is made to develop multiple imputations.

Perhaps the gold standard is the maximum likelihood method, working directly with the likelihood based on observed incomplete data. This option is preferable for a limited set of models for which the software is available. In fact, the multiple imputation based on Bayesian formulation can be viewed as an approximate maximum likelihood method. Specifically, suppose that $L(\theta|D_{obs}, D_{miss})$ is the likelihood that would have been constructed had there been no missing data. The observed data likelihood is,

$$L(\theta|D_{obs}) = \int L(\theta|D_{obs}, D_{miss}) \Pr(D_{miss}|D_{obs})\, dD_{miss} .$$

In the event that the M imputations, $(D_{miss}^{(l)}, l = 1, 2, \ldots, M)$, are draws from the posterior predictive distribution, $\Pr(D_{miss}|D_{obs})$, the observed data likelihood can be approximated by the average,

$$L(\theta|D_{obs}) \approx \frac{1}{M} \sum_{l} L\left(\theta|D_{obs}, D_{miss}^{(l)}\right),$$

of the completed-data likelihoods. That is, the multiple imputation analysis that combines the likelihood-based analysis from each completed data set is approximately equivalent to the analysis based on the observed data likelihood. This discussion is another justification for Bayesian imputation or something very close to it.

Clearly, several possible options exist for a public health researcher to perform a correct analysis with incomplete data. Both user-driven software and commercial software are becoming available to implement these methods. Though most methods rely on a MAR assumption, its lack of applicability is related to the lack of variables that can be used to predict the missing values. Because the missing data are inevitable, a prudent step, from the design perspective, is to investigate potential predictors of variables with missing data and include them in the data-collection process. Such auxiliary variables can include administrative data, neighborhood-level observations, and interviewer observations. These additional variables can be used in the multiple imputation process. It is important that missing data be considered not solely a data analysis problem, but also a design and analysis problem.

APPENDIX

Suppose in a sample of n, r subjects are missing values in x. Let U_R denote the design matrix with r rows and four columns representing intercept, D, E, and $D \times E$. Similarly, let U_M denote the design matrix with n-r rows and four columns for the nonrespondents. Let $\hat{\beta} = (U_R^t U_R)^{-1} U_R^t X_R$ be the least square estimate of the regression of X_R on U_R, where X_R is a vector with r rows containing observed values of the covariate x. Let $s = (X_R - U_R\hat{\beta})^t (X_R - U_R\hat{\beta})$ denote the residual sum of squares. The following algorithm then represents a draw from the posterior predictive distribution of missing covariates conditional on the observed data.

1. Draw a chi-square random variable, c, with r-4 of freedom and define,$\sigma_*^2 = s/c$.
2. Draw r independent standard normal deviates and arrange them as a vector z.
3. Define $\beta_* = \hat{\beta} + \sigma_* T_R z$, where T_R is the square root (Cholesky decomposition) of the matrix $(U_R^t U_R)^{-1}$.
4. Draw n-r independent random normal deviates and arrange them as a vector v.
5. Define $x_* = U_M \beta_* + \sigma_* v$ as imputed values.
6. Repeat steps 1–3 independently to generate multiple imputation.

The *Annual Review of Public Health* is online at http://publhealth.annualreviews.org

LITERATURE CITED

1. Barnard J, Rubin DB. 1999. Small-sample degrees of freedom with multiple imputation. *Biometrika* 86:949–55
2. Binder DA. 1983. On the variances of asymptotically normal estimators from complex survey data. *Int. Statist. Rev.* 51:279–92
3. Carlin JB, Li N, Greenwood P, Coffey C. 2002. Tools for analyzing multiple imputed data sets. Tech. Rep., Univ. Melbourne, Australia
4. David MH, Little RJA, Samuhel ME, Triest RK. 1986. Alternative methods for CPS income imputation. *J. Am. Statist. Assoc.* 81:29–41
5. Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc.* 39:1–38
6. Ford BN. 1983. An overview of hot deck procedures. In *Incomplete Data in Sample Surveys, Vol II: Theory and Annotated Bibliography*, ed. WG Meadow, I Olkin, DB Rubin, pp. 185–206. New York: Academic
7. Glynn RJ, Laird NM. 1986. Regression estimates and missing data: complete case analysis. Tech. Rep., Harvard School of Public Health, Dep. Biostatistics
8. Heckman JI. 1976. The common structure of statistical models of truncation, sample

selection and limited dependent variables, and a simple estimator for such models. *Ann. Econ. Soc. Meas.* 5:475–92

9. Holt D, Elliot D. 1991. Methods of weighting for unit nonresponse (correction: v41, p. 599). *Statistician* 40:333–42
10. Kennickell AB. 1991. Imputation of the 1989 Survey of Consumer Finances: stochastic relaxation and multiple imputation. *Proc. Sec. Surv. Res. Meth. Am. Statist. Assoc.* pp. 1–10
11. Kish L. 1965. *Survey Sampling.* New York: Wiley
12. Kleinbaum DG, Morgenstern H, Kupper LL. 1981. Selection bias in epidemiological studies. *Am. J. Epidem.* 113:452–63
13. Li KH, Meng XL, Raghunathan TE, Rubin DB. 1991. Significance levels from repeated p-values with multiply imputed data. *Statist. Sinica* 1:65–92
14. Li KH, Raghunathan TE, Rubin DB. 1991. Large sample significance levels from multiply-imputed data using moment-based sta tistics and an F-reference distribution. *J. Am. Statist. Assoc.* 86:1065–73
15. Little RJA. 1992. Regression with missing X's: a review. *J. Am. Statist. Assoc.* 87:1227–37
16. Little RJA. 1993. Pattern-mixture models for multivariate incomplete data. *J. Am. Statist. Assoc.* 88:125–34
17. Little RJA. 1995. Modeling the drop-out mechanism in longitudinal studies. *J. Am. Statist. Assoc.* 90:1112–21
18. Little RJA, Raghunathan TE. 1997. Should imputation of missing data condition on all observed variables? *Proc. Sec. Surv. Res. Meth. Am. Statist. Assoc.* pp. 617–22
19. Little RJA, Rubin DB. 2002. *Statistical Analysis with Missing Data.* New York: Wiley
20. Little RJA, Schluchter MD. 1985. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* 72:497–512
21. McKendrick AG. 1926. Applications of mathematics to medical problems. *Proc. Edinburgh Math. Soc.* 44:98–130
22. Raghunathan TE, Grizzle JE. 1995. A split-questionnaire survey design. *J. Am. Statist. Assoc.* 90:55–63
23. Raghunathan TE, Lepkowski JM, van Hoewyk M, Solenberger PW. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodol.* 27:85–95. For associated IVEware software, see http://www.isr.umich.edu/src/smp/ive
24. Raghunathan TE, Siscovick DS. 1996. A multiple imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmacologically treated hypertensive. *Appl. Statist.* 45:335–52
25. Rao JNK. 1996. On variance estimation with imputed survey data. *J. Am. Statist. Assoc.* 91:499–506
26. Rao JNK, Shao J. 1992. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* 79:811–22
27. Rubin DB. 1974. Characterizing the estimation of parameters in incomplete data problems. *J. Am. Statist. Assoc.* 69:467–74
28. Rubin DB. 1976. Inference and missing data (with discussion). *Biometrika* 63:581–92
29. Rubin DB. 1977. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Am. Statist. Assoc.* 72:538–43
30. Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley
31. Rubin DB. 1996. Multiple imputation after 18+ years (with discussion). *J. Am. Statist. Assoc.* 91:473–89
32. Rubin DB, Scehnker N. 1986. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Am. Statist. Assoc.* 81:366–74
33. Rubin DB, Stern H, Vehovar V. 1995.

Handling “don’t know” survey responses: the case of Slovenian plebiscite. *J. Am. Statist. Assoc.* 90:822–28

34. Schafer JL. 1997. *Analysis of Incomplete Multivariate Data*. New York: CRC Press. For associated software, see http://www.stat.psu.edu/∼jls
35. Vach W. 1994. *Logistic Regression with Missing Values in Covariates*. New York: Springer-Verlag
36. Wilks SS. 1932. Moment and distribution of estimates of population parameters from fragmentary samples. *Ann. Math. Stat.* 3:163–95