SHOULD IMPUTATION OF MISSING DATA CONDITION ON ALL OBSERVED VARIABLES?

Roderick Little and Trivellore Raghunathan, University of Michigan
Roderick Little, Biostatistics Department, 1420 Washington Hgts., Ann Arbor MI 48109-2029

Key Words: Multiple Imputation, Nonresponse

1. INTRODUCTION

A common practical approach to missing data is imputation, where missing values are filled in by estimates and the resulting data are analyzed by complete-data methods. Imputation methods until the late 1970's lacked an underlying theoretical rationale. Pragmatic estimates of the missing values were substituted, such as unconditional or conditional means, and inferences based on the filled-in data. The overstatement of precision that results from treating the imputed data as the truth was generally treated as minor and ignored.

Rubin's multiple imputation (MI) theory (Rubin 1977, 1987) put imputation on a firmer theoretical footing, and also provided simple ways of incorporating imputation uncertainty into the inference. Instead of imputing a single set of draws for the missing values, a set of K (say K = 5) datasets are created, each containing different sets of draws of the missing values from their predictive distribution. The analysis of interest is applied to each of the K datasets and results combined in simple ways, such as that discussed in the next section.

MI theory suggests that the imputations in this process should be draws from the predictive distribution of the missing values given the observed data. In particular, suppose we have a dataset with independent observations, and for observation i let $y_{mis,i}$ denote the set of missing variables and let $y_{obs,i}$ denote the set of observed variables. Then imputations should be draws from the predictive distribution of $y_{mis,i}$ given $y_{obs,i}$.. This approach conflicts with widely-held ideas of imputation in some settings. We provide three examples with increasing generality:

Example 1. Missing covariates in regression. Consider the regression of $Y = X_p$ on $X_1,...,X_{p-1}$ for a dataset with Y fully observed and missing values for the covariates $X_1,...,X_{p-1}$. The imputation principle noted above implies that missing covariates in a case are imputed based on the regression of those variables on the observed covariates *and* Y. However many analysts fill in the missing covariates by regressing on the observed covariates, excluding Y. The reasoning is that the inclusion of Y as a predictor of the missing values is circular, given the fact that the filled-in dataset is then used to regress Y on the observed and imputed covariates.

Example 2. Repeated measures with missing data. More generally, many analysts believe imputations should condition only on variables that are exogenous in the system of equations used to model the data. Consider for example a repeated measures problem where a subject is observed at times $T = 1,...,t-1$, is missing at time t, and then reenters the study and is observed at times $t+1,...,T$. From a causal perspective, one might argue that data prior to time t should be used to impute data at time t but not data after time t. However the MI principles above imply that observed data values at times $t+1,...,T$ should be used to help impute the values of missing values at time t, despite the reverse causal direction.

Example 3. Imputations for analyses involving different but overlapping sets of variables. Data bases are generally subject to a broad range of analyses involving different sets of variables. Suppose that analysis (A) of a dataset involves a set of variables S_A and analysis (B) involves a set of variables S_B, and these two sets both include a variable X that has values missing. Constrast the following three analysis strategies:

(1) Carry out analysis (A) based on S_A and analysis (B) based on S_B using methods that do not involve imputation of the missing values of X, such as maximum likelihood applied to the set of variables involved.

(2) Carry out analysis (A) with imputations of X based on the data available in S_A, and analysis (B) with imputations of X based on data available in S_B.

(3) Carry out analyses (A) and (B) with imputations of X based on the combined information in S_A and S_B.

Analysis (1) and (2) are common choices, but MI theory suggests that (3) is the best analysis, since it conditions on the available information about X.

In this paper we reaffirm the rationale for imputation that conditions on all the available data in these settings. However, we also show that for certain nonignorable nonresponse models the imputation principle described above can lead to methods that do not use information on all the available variables.

2. MULTIPLE IMPUTATION THEORY

The theory of MI is model-based, and is founded on a simulation approximation of the Bayesian posterior distribution of the parameters given the observed data.

Specifically, let $X = (x_{ij})$ represent an $n \times p$ data matrix, let M be an $(n \times p)$ missing-data indicator matrix with entries $M_{ij} = 1$ if x_{ij} is missing and $M_{ij} = 0$ is x_{ij} is observed. Let X_{obs} be the observed data and X_{mis} the missing components of X. Let $p(X, M|\theta)$ be the distribution of X and M indexed by model parameters θ, and let $p(\theta)$ be a prior distribution for θ. The posterior distribution for θ given the observed data (X_{obs}, M) is related to the posterior distribution given hypothetical complete data (X, M) by the expression:

$$p(\theta|X_{obs}, M) = \int p(\theta|X, M) p(X_{mis}|X_{obs}) dX_{mis}$$

MI approximates this expression as:

$$p(\theta|X_{obs}, M) \cong \frac{1}{K}\sum_{k=1}^{K} p(\theta|X^{(k)}, M), \qquad (1)$$

where $X^{(k)} = (X_{obs}, X_{mis}^{(k)})$ is an imputed data set with missing values filled in by a draw from the posterior predictive distribution of the missing data X_{mis} given the observed data X_{obs}. and M:

$$X_{mis}^{(k)} \sim p(X_{mis}|X_{obs}, M), \qquad (2)$$

and $p(\theta|X^{(k)}, M)$ is the posterior for θ based on the filled-in data set $X^{(k)}$. The posterior mean and covariance matrix of θ can be approximated similarly as:

$$E(\theta|X_{obs}, M) \cong \bar{\theta} \equiv \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}^{(k)}, \ \hat{\theta}^{(k)} = E(\theta|X^{(k)}, M) \, (3)$$

$$Var(\theta|X_{obs}, M) \cong \frac{1}{K}\sum_{k=1}^{K} Var(\theta|X^{(k)}, M) + \frac{K+1}{K}\left(\frac{1}{K-1}\sum_{k=1}^{K}(\hat{\theta}^{(k)} - \bar{\theta})(\hat{\theta}^{(k)} - \bar{\theta})^T\right) \qquad (4)$$

These expressions form the basis for MI inference of the filled-in data sets. Eq. (3) indicates that the combined estimate of a parameter is obtained by averaging the estimates from each of the filled-in data sets. Eq. (4) indicates that the covariance matrix of the estimate is obtained by averaging the covariance matrix from the filled-in data sets and adding the sample covariance matrix of the estimates $\hat{\theta}^{(k)}$ from each of the filled-in data sets, which captures the added uncertainty from imputation that is missed by single-imputation methods; the $(K+1)/K$ factor is a small-sample correction that improves the approximation.

The following comments on this approach are germane to our later discussion:

(A) Although the above MI theory is Bayesian, simulations show that inferences based on (1)-(4) have good frequentist properties, at least if the model and prior are reasonable. For discussion of the properties of MI under model misspecification, see for example Rubin (1996), Fay (1996) and Rao (1996) and the literature cited in those papers.

(B) An important feature of the method is that *draws* of the missing values are imputed rather than *means*. Means would be preferable if the objective was to obtain the best estimates of the missing values, but mean imputation has drawbacks when the objective is to make inferences about parameters. Imputation of draws entails some loss of efficiency for point estimation, but this loss is considerably reduced by the averaging over the K multiply-imputed data sets in (3). The gain from imputing draws is that it yields valid inferences for a wide range of estimands, including nonlinear functions such as percentiles and variances (Little 1988).

(C) The difficulty in implementing MI is in obtaining draws from the posterior distribution of X_{mis} given X_{obs}, which often has an intractable form. Since draws from the posterior distribution of X_{mis} given X_{obs} and θ are often much easier, a simpler scheme is to draw from the posterior distribution of X_{mis} given X_{obs} and $\tilde{\theta}$, where $\tilde{\theta}$ is an easily computed estimate of θ such as that obtained from the complete cases. This approach ignores uncertainty in estimating θ, and is termed *improper* in Rubin (1987). It yields acceptable approximations when the fraction of missing data is modest, but leads to overstatement of precision with large amounts of missing data. In the latter situations one option is to draw $\theta^{(k)}$ from its asymptotic distribution and then impute X_{mis} from its posterior distribution given X_{obs} and $\theta^{(k)}$. A better but more computationally-intensive approach is to cycle between draws $X_{mis}^{(t)} \sim P(X_{mis}|X_{obs}, \theta^{(t-1)})$ and $\theta^{(t)} \sim p(\theta|X_{obs}, X_{mis}^{(t)})$, an application of the Gibbs' sampler (Tanner and Wong 1987).

(D) The formulation above requires specification of the joint distribution of X and M. Fixed covariates that are fully observed in the data set can be treated as fixed in the analysis and do not need to be modeled. In particular in Example 1, a distribution only needs to be specified for the joint distribution of the outcome and incomplete covariates conditional on the fully observed covariates. A more important point is that if the missing data are missing at random (MAR) in that the distribution of M given X depends on the values of observed variables X_{obs}, then inference can be based on a model for X alone rather than on a model for the joint distribution of X and M (Rubin 1976: Little and Rubin 1987). Specifically Eqs. (1) -- (4) can replaced by (1I) -- (4I):

$$p(\theta|X_{obs}) \cong \frac{1}{K}\sum_{k=1}^{K} p(\theta|X^{(k)}), \qquad (1I)$$

$$X^{(k)} = (X_{obs}, X_{mis}^{(k)}),\ X_{mis}^{(k)} \sim p(X_{mis}|X_{obs}), \qquad (2I)$$

$$E(\theta|X_{obs}) \cong \bar{\theta} \equiv \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}^{(k)},\ \hat{\theta}^{(k)} = E(\theta|X^{(k)}) \qquad (3I)$$

$$Var(\theta|X_{obs}) \cong \frac{1}{K}\sum_{k=1}^{K} Var(\theta|X^{(k)}) + \frac{K+1}{K}\left(\frac{1}{K-1}\sum_{k=1}^{K}(\hat{\theta}^{(k)}-\bar{\theta})(\hat{\theta}^{(k)}-\bar{\theta})^{T}\right) \qquad (4I)$$

where the conditioning on M has been dropped. This approach is called inference *ignoring the missing-data mechanism*, and is attractive since in practice modeling the missing-data mechanism is difficult, and results are vulnerable to model misspecification.

(E) Another important feature of MI is that it involves draws from the conditional distribution of the missing data given all the observed data. This feature contrasts with other approaches to imputation that do not condition on all the observed variables in the data set. We now examine this issue further using the examples of regression and longitudinal data analysis mentioned in the introduction.

Example 1. Missing covariates in regression (contd.) When imputing missing covariates in regression, it is common practice not to condition on the values of observed outcome variable Specifically, consider a linear regression of $Y \equiv X_p$ on covariates $X_1,\ldots,X_{p-1}$, when $X_1,\ldots,X_{p-2}$ and Y are fully observed and X_{p-1} is missing for some observations. MI theory dictates that missing values of X_{p-1} should be imputed based on the conditional distribution of X_{p-1} given $X_1,\ldots,X_{p-2}$ and Y. However it is commonly considered circular to use Y for imputation in this setting, and imputation is instead based on the conditional distribution of X_{p-1} given $X_1,\ldots,X_{p-2}$. In particular, one strategy is to regress X_{p-1} on $X_1,\ldots,X_{p-2}$ using the complete cases, and then fill in missing values of X_{p-1} with predictions from this regression. Analysis of the filled-in data set yields valid estimates of the regression of Y on $X_1,\ldots,X_{p-2}$, provided missingness does not depend on X_{p-1} or Y.

The key to whether imputations should condition on Y is whether a draw or a mean is imputed. If a draw is imputed, as in MI, then bias results if the imputations do *not* condition on Y -- specifically, if X_{p-1} is drawn from the conditional distribution of X_{p-1} given $X_1,\ldots,X_{p-2}$ then the regression coefficient of Y on X_{p-1} computed from the filled-in data is attenuated. If on the other hand an estimated conditional mean is imputed, then bias results if imputations condition on Y. In fact, the E-Step of the EM algorithm for maximum likelihood based on a normal model with ignorable missing data does impute the mean of X_{p-1} conditional on $X_1,\ldots,X_{p-2}$ and Y, but the algorithm incorporates a correction to the covariance matrix of the variables that adjusts for the bias (Beale and Little 1975). Imputing the conditional mean of X_{p-1} given $X_1,\ldots,X_{p-2}$ yields consistent estimates for the regression on the filled-in data, but the resulting estimates involve a loss of efficiency, and can in fact be less efficient than estimates based on the complete cases (Gourieroux and Montfort 1981, Conniffe 1983a,b, Little 1992). As discussed in these papers, the efficiency of the method can be increased by down-weighting the imputed cases, but obtaining optimal weights and valid standard errors is not straightforward, particularly for more complex patterns of missing data.

To see why the circularity of the MI approach is not a problem, consider its most rigorous implementation using the Gibbs' sampler. The latter appears circular since draws of X_{mis} condition on θ and draws of θ condition on X_{mis}. Nevertheless, the Gibbs' cycle iterates to a draw from the joint posterior of θ and X_{mis}, and hence yields Bayesian inferences about θ that have optimal large-sample properties (including consistency) from a frequentist perspective. MI is attractive since it allows for imputation uncertainty and is asymptotically efficient as the number K of multiple imputes tends to infinity. High efficiency can be obtained with a small value of K (say 5) if the fraction of incomplete values is fairly small.

Example 2. Repeated measures with missing data (contd.) Consider a repeated measures problem where a variable X is initially observed at two times $T = 1, 2$. Under a simple causal model, interest lies in the regression of X_2, the value of X at time 2 on X_1, the value of X at time 1. Suppose that X_1 is fully observed but values of X_2 are missing for some cases not recorded at $T = 2$. If the data are MAR and parameters of the marginal distribution of X_1 and the conditional distribution of X_2 given X_1 are distinct, then it is well known that the cases with X_2 missing provide no information about the regression of X_2 on X_1 and can be discarded. Indeed MI inference that imputes the missing values of X_2 is asymptotically equivalent to complete-case analysis as K becomes large, and there is no gain from including the incomplete cases in the analysis.

Now suppose that another wave of data is collected, and values of X at time 3, say X_3, are recorded. Some fraction of the cases missing at time 2 reenter the sample, and hence have values of X_1 and X_3 recorded. The question is whether data from this wave provide information for the regression of X_2 on X_1; from the imputation point of view this depends on whether imputations of missing values of X_2 are allowed to condition on the values of X_3. A perspective that does not allow imputes to condition on endogenous variables would not allow imputes of X_2 to condition on X_3 for the regression of X_2 on X_1, since X_3 is not exogenous to this regression if causality follows the direction of time. This implies that for the purposes of this regression imputes of X_2 should not condition of X_3, and data on X_3 should be discarded. However, intuitively the data on X_3 could be very useful in predicting missing values at time 2; indeed in the extreme case where X_2 and X_3 are very highly correlated, an imputation method that conditions on X_1 and X_3 can essentially recover all the missing information for the regression for cases with X_1 and X_3 observed and X_2 missing.

MI recovers this information by imputing draws of X_2 from the conditional distribution given X_1 and X_3 for cases that reenter the sample at time 3. The important point is that this form of MI provides valid inference about any parameters of the joint distribution of X_2 and X_3 given X_1 under the assumed model, and hence in particular yields valid inferences for the parameters of the regression of X_2 on X_1. The lack of exogeneity of X_3 does not affect the validity of MI inference for the parameters that are of causal interest.

Example 3. Imputations for analyses involving different but overlapping sets of variables (contd.)

MI theory leads to Strategy (3), where imputations are based on the combined set of variables S_A and S_B. The imputations are potential improved by the inclusion of a larger set of predictors, the MAR assumption is improved in situations where missingness depends on variables in both sets, and consistency of the two analyses is improved because the treatment of missing values in the same for both analyses. Compared with Strategy (1) where missing values are not imputed, MI has the useful feature that variables not in the main analysis model can be readily included in the imputation model.

3. CAVEATS TO CONDITIONING ON ALL OBSERVED VARIABLES.

The MI approach requires a suitable model for the predictive distribution of the missing data given the observed data. Ideally, the imputation model should not make strong assumptions that might conflict with the model for the main analysis. On the other hand the number of parameters in the model needs to be tailored to the size of the data set at hand. For example a standard regression model with noninformative priors is not appropriate when the number of conditioning variables for imputation exceeds the number of observations. Informative priors, or more pragmatic approaches such as ridge regression or variable subset selection may be needed to reduce the size of the model in such cases.

As noted in Section 1, when the data are not MAR, a model for joint distribution of X and the missing-data indicator matrix M is needed. *Selection models* specify this distribution as:

$$f(X, M|\gamma, \psi) = f(Y|\gamma) f(M|Y, \psi), \qquad (5)$$

where $f(Y|\gamma)$ is the model in the absence of missing values, $f(M|Y,\psi)$ is the model for the missing-data mechanism, and γ and ψ are unknown parameters. In contrast, *pattern-mixture models* specify:

$$f(Y, M|\pi, \phi) = f(Y|M, \phi) f(M|\pi), \qquad (6)$$

where ϕ and π are unknown parameters and now the distribution of Y is conditioned on the missing-data pattern M (Little and Rubin 1987; Little 1993).

Most of the literature on missing data has concerned selection models of the form (5), which are natural when interest concerns parameters γ of the complete-data distribution. Important examples are probit (Heckman 1976; Amemiya 1984), and logit (Greenlees, Reece and Zieschang 1982; Diggle and Kenward 1994). selection models. Inferences about γ and ψ are dangerously unstable and sensitive to misspecification of the model for the missing-data mechanism (Little 1985; Little and Rubin 1987, Chapter 11; Glynn, Laird and Rubin 1993; Stolzenberg and Relles 1990).

Pattern-mixture models seem more natural when missingness defines a distinct stratum of the population of intrinsic interest, such as individuals reporting "don't know" in an opinion survey. However, they can also provide inferences for parameters γ of the complete-data distribution, by expressing the parameters of interest as functions of the pattern-mixture model parameters ϕ and π. Pattern-mixture models often yield inestimable parameters, but can avoid the need to model the form of the missing-data mechanism explicitly, since the mechanism is reflected in restrictions on the model parameters.

We believe that when the nonresponse mechanism is not MAR, in most cases imputation based on an ignorable model that conditions on all available data will tend to reduce nonresponse bias compared to alternative methods, but this is not always the case. An ignorable model that fails to condition on all the observed variables may be closer to the true nonignorable model than an ignorable model that conditions on all the observed variables, and hence yield superior imputations. The following example from Little (1994) illustrates this situation.

Example 3. Nonignorable models for longitudinal data with two time points.

Consider as in Example 2 a repeated measures problem where x_{ij} is the value of a variable X for subject i measured at time j, $j = 1, 2$, and $\{x_{i1}\}$ are fully observed but values of $\{x_{i2}\}$ are missing for some cases not recorded at time 2. Let $m_i = 1$ if x_{i2} is missing, $m_i = 0$ if x_{i2} is present, and consider the following pattern-mixture model:

(a) given $m_i = r$, $x_i = (x_{i1}, x_{i2})^T$ is iid bivariate normal with mean and covariance matrix $\phi^{(r)} = \{\mu^{(r)}, \Sigma^{(r)}\}$;

(b) m_i is marginally iid Bernoulli with $pr(m_i = 1) = \pi$.

(c) $p(m_i = 1 | x_{i1}, x_{i2}, \lambda)$ = $g(x_{i1} + \lambda x_{i2})$ for arbitrary function g and known λ .

The distributions in (a) and (b) together contain eleven parameters, of which eight are estimable from the data and three, corresponding to the conditional distribution of X_2 given X_1 for the incomplete cases, are not identified. Assumption (c) implies that

$$p(X_1 | X_2^*, M = 1, \phi) = p(X_1 | X_2^*, M = 0, \phi),$$

where $X_2^* = X_1 + \lambda X_2$, which yields three restrictions on the parameters, namely equality of the intercepts, slopes and residual variances of the conditional distribution of X_1 given $X_1 + \lambda X_2$ for complete and incomplete cases. These constraints just identify the parameters $\phi^{(r)}$ of the model. The coefficient λ measures the extent to which missingness depends on X_2 rather than X_1. In particular if $\lambda = 0$ then missingness depends on X_1 and is MAR, if $\lambda = \infty$ then missingness depends entirely on X_2, and if $\lambda = -1$ then missingness depends on the difference $X_2 - X_1$. We would like to be able to estimate λ simultaneously with the other model parameters, but the data supply no information about λ (Little 1994). Hence results need to be based on an *a priori* choice of λ, or a sensitivity analysis for a range of values of λ.

The ML estimates of the parameters of this model are derived in Little (1994). In particular the ML estimates of the mean and variance of X_2 and covariance of X_1 and X_2 combined over patterns are:

$$\hat{\mu}_2 = \bar{x}_2 + b_{21\cdot1}^{(\lambda)}(\hat{\mu}_1 - \bar{x}_1) \tag{7}$$

$$\hat{\sigma}_{22} = s_{22} + (b_{21\cdot1}^{(\lambda)})^2(\hat{\sigma}_{11} - s_{11}) \tag{8}$$

$$\hat{\sigma}_{12} = s_{12} + b_{21\cdot1}^{(\lambda)}(\hat{\sigma}_{11} - s_{11}) \tag{9}$$

where

$$b_{21\cdot1}^{\lambda} = \frac{\lambda s_{22} + s_{12}}{\lambda s_{12} + s_{11}}. \tag{10}$$

Here $\bar{x}_1, \bar{x}_2, s_{11}, s_{22}, s_{12}$ are the sample means, variances and covariance of X_1 and X_2 for the complete cases, and $\hat{\mu}_1$ and $\hat{\sigma}_{11}$ is the sample mean and variance of X_1 over complete and incomplete cases. When λ = 0, the missing data are MAR, and these estimates correspond to imputation of missing values of X_2 based on the regression of X_2 on X_1 estimated from the complete cases. However if λ = $\tilde{\lambda} \equiv -\beta_{12.2}^{(0)}$, then substituting the ML estimate of λ in (10) yields $b_{21\cdot1}^{(\hat{\lambda})} = 0$, and (7) - (9) reduce to complete-case estimates. These correspond to the imputation of X_2 based on its marginal distribution, without conditioning on values of X_1 for the incomplete cases. The implication is that if $\lambda = \tilde{\lambda}$ is thought to be more plausible than λ = 0, then imputations that fail to condition on the values of X_1 are preferable to imputations that do! This conclusion may seem to contradict statements in Section 2, imputations under the nonignorable model with $\lambda = \tilde{\lambda}$ still condition on observed values of X_1, but the coefficient $b_{21\cdot1}^{(\tilde{\lambda})}$ is zero so that in effect values of X_1 are ignored.

4. CONCLUSIONS

1. MI provides a general theoretical principle for filling in missing values and propagating the effects of imputation error on inferences. MI theory dictates that imputations should condition on all available information in a case, including variables that are not included in a particular analysis or that are causally endogenous to the missing variables.

2. Ignorable missing data models are convenient in that they avoid the need to specify the missing-data mechanism, and they are good representation of reality when good covariates that characterize nonrespondents are available.

3. When nonignorable nonresponse mechanisms are thought to operate, nonignorable nonresponse models may play a role in the analysis, particularly as part of a sensitivity analysis. Particular nonignorable nonresponse models can yield imputations that correspond to those from ignorable models that do not condition on the all available data. If such nonignorable models are better descriptions of reality than the ignorable model, then failing to condition on particular covariates can be justified.

ACKNOWLEDGEMENTS

This research was supported by Grant DMS 9408837 from the National Science Foundation.

REFERENCES

Amemiya, T. (1984). Tobit Models: a Survey. *Journal of Econometrics,* 24: 3-61.

Beale, E.M.L., and Little, R.J.A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society B,* 37: 129-145.

Conniffe, D. (1983a). Comments on the weighted regression approach to missing values, *Economic and Social Review*, 14, 259-272.

Conniffe, D. (1983b). Small-sample properties of estimators of regression coefficients given a common pattern of missing data. *Review of Economic Studies*, L, 111-120.

Diggle, P. and Kenward, M.G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, 43, 49-94.

Fay, R.E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. *Journal of the American Statistical Association*, 91, 490-498.

Glynn, R. Laird, N.M. and Rubin, D.B. (1993). Multiple imputation in mixture models for nonignorable nonresponse with follow-ups. *Journal of the American Statistical Association,* 88, 984-993.

Gourieroux, C. and Montfort, A. (1981). On the problem of missing data in linear models. *Review of Economic Studies*, XLVIII, 579-586.

Greenlees, W.S., Reece, J.S. and Zieschang, K.D. (1982). Imputation of missing values when the probability of nonresponse depends on the variable being imputed. *Journal of the American Statistical Association,* 77, 251-261.

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

Little, R.J.A. (1985). A note about models for selectivity bias. *Econometrica,* 53, 1469-1474.

Little, R.J.A. (1988). Missing data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6: 287-301.

Little, R.J.A. (1992). Regression with incomplete *X*'s; a review. *Journal of the American Statistical Association*, 87: 1227-1237.

Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association,* 88, 125-134.

Little, R.J.A. (1994). A class of pattern-mixture models for normal missing data. *Biometrika,* 81, 471-483.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York: John Wiley.

Rao, J.N.K. (1996). On variance estimation with imputed data. *Journal of the American Statistical Association*, 91: 499-506.

Rubin, D.B. (1976). Inference and missing data. *Biometrika,* 63, 581-592.

Rubin, D.B. (1977). Multiple imputations in sample surveys -- a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section, American Statistical Association 1977*, 20-34.

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

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

Stolzenberg, R.M., and Relles, D.A. (1990). Theory testing in a world of constrained research design - the significance of Heckman's censored sampling bias correction for nonexperimental research. *Sociological Methods and Research*, 18: 395-415.

Tanner, M.A., and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82: 528-550.