Not Asked and Not Answered: Multiple Imputation for Multiple Surveys

Andrew GELMAN, Gary KING, and Chuanhai LIU

We present a method of analyzing a series of independent cross-sectional surveys in which some questions are not answered in some surveys and some respondents do not answer some of the questions posed. The method is also applicable to a single survey in which different questions are asked or different sampling methods are used in different strata or clusters. Our method involves multiply imputing the missing items and questions by adding to existing methods of imputation designed for single surveys a hierarchical regression model that allows covariates at the individual and survey levels. Information from survey weights is exploited by including in the analysis the variables on which the weights were based, and then reweighting individual responses (observed and imputed) to estimate population quantities. We also develop diagnostics for checking the fit of the imputation model based on comparing imputed data to nonimputed data. We illustrate with the example that motivated this project: a study of pre-election public opinion polls in which not all the questions of interest are asked in all the surveys, so that it is infeasible to impute within each survey separately.

KEY WORDS: Bayesian inference; Cluster sampling; Diagnostics; Hierarchical models; Ignorable nonresponse; Missing data; Political science; Sample surveys; Stratified sampling.

1. INTRODUCTION

Multiple imputation is a general approach for handling nonresponse in sample surveys. In particular, it is often useful to use automatic methods, based on fitting saturated or nearly saturated models, to impute missing data, with the understanding that once the imputations have been obtained, later users can analyze the completed datasets as they see fit (see Belin et al. 1993; Meng 1994; Rubin 1987, 1996). (Also see Fay 1996 and Rao 1996 for critical perspectives on multiple imputation.) Algorithms are available and in use for imputing missing data in a single sample survey based on normal (Liu 1993; Rubin and Schafer 1990; Schafer 1997) and t (Liu 1995) distributions and the general location model (Liu and Rubin 1998; Schafer 1997).

When imputing missing data from several sample surveys, there are two obvious ways to use existing single-survey methods: (1) separately imputing the missing data from each survey or (2) combining the data from all of the surveys and imputing the missing data in the combined "data matrix." Both of these methods have problems. The first approach is difficult if there is a large amount of missingness in each individual survey. For example, if a particular question is not asked in one survey, then there is no general way to impute it without using information from other surveys or some additional knowledge about the relation between responses to that question and to other questions asked in the survey. The second method does not account for differences between the surveys—for example, if they are conducted at different times, use different sampling methodologies, or are conducted by different survey organizations.

Our approach is to compromise by fitting a separate imputation model for each survey, but with the parameters in the different surveys linked with a hierarchical model. This method should have the effect that imputations of item nonresponse in a survey will be determined largely by the data from that survey, whereas imputations for questions not asked in a survey will be determined by data from the other surveys in the population as well as by available responses to other questions in that survey. This effect of partial pooling, with the amount of pooling depending on the amount of available data, is typical of Bayesian inference in hierarchical models or meta-analysis (see, e.g., Belin et al. 1993; DuMouchel and Harris 1983; Efron and Morris 1975; Gatsonis, Normand, Morris, and Liu 1992; Rubin 1980). The hierarchical regression structure also allows us to include covariates both at the individual and survey levels. (For an approach to hierarchical regression using econometric methods, see Franklin 1989.) A related Bayesian approach to the problem of missing covariates in a regression analysis of cross-sectional surveys has been given by Dominici et al. (1996).

Another relevant area of application is stratified and cluster sampling. Appropriate analysis of sample surveys includes information used in the design, including stratification and clustering. (For perspectives from survey sampling practice, Bayesian inference, and multiple-imputation inference, see Gelman, Carlin, Stern, and Rubin 1995; Kish 1965; Rubin 1996) If strata or clusters are expected to differ in their mean responses (as will generally be the case), then it would be reasonable to apply a hierarchical model instead of imputing using a common distribution for all the respondents irrespective of stratum/cluster. For cluster

Andrew Gelman is Associate Professor, Department of Statistics, Columbia University, New York, NY, 10027 (E-mail: gelman@wald.stat.columbia.edu), (http://www.stat.columbia.edu/~gelman/). Gary King is Professor, Department of Government, Harvard University, Cambridge, MA 02138 (E-mail: king@harvard.edu), (http://gking.harvard.edu). Chuanhai Liu is technical staff member, Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 (E-mail: liu@research.bell-labs.com). The authors thank Xiao-Li Meng, Donald Rubin, Alan Zaslavsky, and several reviewers for helpful comments; the National Science Foundation for grants SBR-9223637, SBR-9321212, DMS-9404305, and Young Investigator grant DMS-9457824; and the Research Council of Katholieke Universiteit Leuven for fellowship F/96/9. The data and computer code necessary to replicate the results in this article will be deposited in the STATLIB data archive.

sampling, the hierarchical model has the additional advantage of immediately generalizing to the unsampled clusters. Our method might be particularly appropriate to surveys in which different questions are asked to respondents in different strata (see Raghunathan and Grizzle 1995).

In this article we present a specific method for extending a standard multiple imputation algorithm based on multivariate normal models. We illustrate with the example that motivated this work, a study of 51 public opinion polls preceding the 1988 U.S. Presidential election. In the presentation of the example, we discuss some practical issues in using the imputations, including concerns about discrete data and accounting for survey weights in the imputation and analysis of results. In presenting the results for the example, we illustrate some novel graphical methods for summarizing the results of the multiple imputations and checking the fit of the imputation model and the calibration of the between-imputation variability.

2. THE MODEL

2.1 Notation and Basic Assumptions

Suppose that S sample surveys are conducted and we are analyzing Q questions, each of which is asked in at least one of the S surveys. (Equivalently, S could be the number of strata or clusters within a single survey; for simplicity, we work with the multiple-survey context here.) When any of the Q questions is not asked in some surveys, we imagine that it could have been asked but all of the responses to this question are missing. In addition, there can be item nonresponse, so not all the survey respondents respond to every question asked of them. To handle both situations, we augment the data such that the complete data consist of the same Q questions in all of the S surveys. We denote by $y_{s,i} = (y_{s,i,1}, \ldots, y_{s,i,\mathrm{Q}})'$ the responses of individual i in survey s to all of the Q questions. Some of the elements of $y_{s,i}$ may be missing. Letting N_s be the number of the respondents in survey s, the (partially unobserved) complete data have the form

$$\{\{(y_{s,i,1}, \ldots, y_{s,i,\mathrm{Q}})' \colon i = 1, \ldots, N_s\} \colon s = 1, \ldots, S\}. \quad (1)$$

We assume that the data are *missing at random*; that is, the probability of missingness depends only on observed data included in the model (Rubin 1976). This is a reasonable assumption here, because almost all of the missingness is due to unasked questions. If clear violations of missingness at random occur (e.g., a question about defense policy may be more likely to be asked when the country is at war), then additional survey-level variables should be included in the model until missingness at random is once again a reasonable assumption (e.g., including a variable for the level of international tension).

We further assume that the rate of missingness provides no information about the underlying responses. That is, we assume that the parameters of the missing-data process are distinct from the parameters of the data model, so that the missing-data mechanism is *ignorable* (see Rubin 1976).

2.2 Hierarchical model

The simplest model for imputing the missing values in the data in (1) that make use of the data structure of the multiple surveys is multivariate normal at the individual level with mean vector μ_s for survey s and a common variance matrix Ψ,

$$y_{s,i}|(\mu_s, \Psi, \theta, \Sigma) \overset{\text{ind}}{\sim} N_{\mathrm{Q}}(\mu_s, \Psi) \quad (i = 1, \ldots, N_s;\ s = 1, \ldots, S), \quad (2)$$

with the means exchangeable at the survey level,

$$\mu_s|(\theta, \Sigma) \overset{\text{ind}}{\sim} N_{\mathrm{Q}}(\theta, \Sigma) \qquad (s = 1, \ldots, S), \quad (3)$$

where θ is a vector of means and Σ is a $(Q \times Q)$ diagonal matrix, diag $(\sigma_1^2, \ldots, \sigma_{\mathrm{Q}}^2)$. The model in equations (2) and (3) allows for pooling information from all the S surveys and imputing all of the missing values, including those to the questions not asked in some of the surveys. The effect of the pooling is illustrated with the example in Section 4. The example also suggests that factors at the survey level, such as organization effects and time trend, should be included in the model for multiple imputation. The assumption that the variance matrix Ψ is the same for all surveys could be tested by, for example, dividing the surveys nonrandomly into two groups (for example, early surveys and late surveys) and estimating separate matrices Φ for the two groups.

Now suppose that we have data on P variables of interest at the survey level. Let $x = (x_1, \ldots, x_P)'$ be the vector of the P survey-level covariates. We assume that x is fully observed for each of the S surveys. We denote by x_s the fully observed P covariates of the sth survey and write $X = \{x_s \colon s = 1, \ldots, S\}$. We consider the following hierarchical model:

$$y_{s,i}|(\mu_s, \Psi, X, \beta, \Sigma) \overset{\text{ind}}{\sim} N_{\mathrm{Q}}(\mu_s, \Psi) \quad (i = 1, \ldots, N_s;\ s = 1, \ldots, S), \quad (4)$$

$$\mu_s|(X, \beta, \Sigma) \overset{\text{ind}}{\sim} N_{\mathrm{Q}}(\beta x_s, \Sigma) \qquad (s = 1, \ldots, S), \quad (5)$$

where β is the $(Q \times P)$ matrix of the regression coefficients of μ on x. Because Σ is diagonal, (5) represents Q linear regression models with normal errors,

$$\mu_{s,j}|(X, \beta, \Sigma) \sim \mathrm{N}(x_s'\beta_j', \sigma_j^2) \quad (s = 1, \ldots, S;\ j = 1, \ldots, Q), \quad (6)$$

where $\mu_{s,j}$ is the jth component of μ_s and β_j is the jth row of β.

Following Liu (1993), we use the following noninformative prior distribution for (Ψ, β, Σ):

$$p(\Psi, \beta, \Sigma) = p(\Psi)p(\beta)\prod_{q=1}^{Q} p(\sigma_q^2) \propto |\Psi|^{-(\mathrm{Q}+1)/2}. \quad (7)$$

If there are fewer than Q completely observed units (i.e., individuals with responses on all the questions), then it is necessary to use a proper prior distribution for Ψ. A minimally informative conjugate prior density when there are

no completely observed units (as would happen, for example, if there were no survey in which all Q questions were asked) is inverse-Wishart with $\nu = Q$ df; that is, $p(\Psi) \propto |\Psi|^{-(\nu+Q+1)/2} \exp(-\frac{1}{2}\text{tr}(\Psi_0\Psi^{-1}))$, where Ψ_0 is a positive-definite "prior estimate" of Ψ. In realistic examples with moderate or large sample sizes and nondegenerate missing-data patterns, this prior distribution will be essentially irrelevant (except for serving the mathematical function of ensuring a proper posterior distribution). For the prior scale matrix Ψ_0, one can use a rough approximation such as a diagonal matrix with elements set to the marginal variances of the Q outcomes. If this proper prior distribution is used, then it is to be treated as ν additional data points when updating Ψ in the subsequent computations. It can also be appropriate to use a proper prior distribution for Σ.

3. COMPUTATION

The model in (2) and (3) is computationally a special case of that in (4) and (5). Here we describe a method to impute the missing values in data (1) under the model in (4) and (5). Our method, which is an extension of that of Schafer (1997), uses two basic steps: data augmentation to form a monotone missing-data pattern and the Gibbs sampler to draw simulations from the joint posterior distribution of the missing data and parameters. We go beyond the work of Schafer (1997) in adapting this method to a hierarchical data structure that includes information at the individual and survey levels.

For incomplete multivariate normal data, Rubin and Schafer (1990) proposed a data augmentation scheme called monotone data augmentation (MDA) for efficiently creating multiple imputations (Liu 1993, 1995, 1996; Schafer 1997). A rectangular dataset $\{(y_{i,1}, \ldots, y_{i,m}): i = 1, \ldots, N\}$ with missing values is said to have a monotone pattern if the data can be sorted in such a way that $y_{i,j}$ is observed if $y_{i+1,j}$ is observed for $j = 1, \ldots, m$ and $i = 1, \ldots, n-1$. MDA is the algorithm that applies the data augmentation algorithm (Tanner and Wong 1987) to a (complete) monotone-pattern dataset, which is created by including those missing values that destroy the monotone pattern. MDA promises fast converging iterative simulation methods by disregarding the missing values of a monotone pattern during iterative simulations. After MDA converges, all of the missing values in the rectangular dataset can be imputed to create a complete rectangular dataset. We use a method-of-moments estimate from fully observed units as a starting point for the iterative data augmentation algorithm.

MDA is very effective in multiple imputation for multiple surveys, because the data can be sorted so that a large portion of the missing values fall into a monotone pattern due to the fact that some questions are not asked in some surveys. First, we sort the data consisting of the S datasets from the S surveys so that a portion of the missing values fall into a monotone pattern, which has Q possible observed-data (or missing-data) patterns. The resulting data matrix can be described as

$$y_{\text{mp}} = \{(y^{(k)}_{s_i,i,k}, \ldots, y^{(k)}_{s_i,i,\text{Q}}): \quad i = 1, \ldots, n_k;\ k = 1, \ldots, Q\}, \quad (8)$$

where k indexes the observed-data pattern and s_i represents the survey containing the ith respondent in the kth observed-data pattern. The data in (8) may still contain missing values. We denote the set of all of the missing values in (8) by $y_{\text{mp,mis}}$ and the set of all the observed values by y_{obs}. Thus we have $y_{\text{mp}} = \{y_{\text{obs}}, y_{\text{mp,mis}}\}$. Figure 1 illustrates a constructed monotone data pattern for the preelection surveys, with the variables arranged in decreasing order of proportion of missing data.

We use the Gibbs sampler (Gelfand and Smith 1990; Geman and Geman 1994; Tanner and Wong 1987), an iterative algorithm for obtaining draws of a set of m variables $\xi_1, \ldots, \xi_m$ from their joint distribution. Each iteration of the Gibbs sampler consists of a sequence of steps, and each takes a draw of a subset of $\{\xi_1, \ldots, \xi_m\}$ from their conditional distribution given the remaining variables in $\{\xi_1, \ldots, \xi_m\}$ with each of the conditioning variables fixed at its most current draw. Under mild conditions, the distribution of the Gibbs sequence will converge to the joint distribution of $(\xi_1, \ldots, \xi_m)$ if each of the $\xi_1, \ldots, \xi_m$ is visited infinitely often.

Using the data augmentation scheme in (8) and the model defined by (4), (5), and (7), we have the observed data y_{obs} and all of the unknowns $\{y_{\text{mp,mis}}, \Psi, \mu_1, \ldots, \mu_S, \beta, \Sigma\}$. To take draws of $\{y_{\text{mp,mis}}, \Psi, \mu_1, \ldots, \mu_S, \beta, \Sigma\}$ from their posterior/predictive distribution given the observed values y_{obs}, we use the version of the *monotone* Gibbs sampler where each iteration consists of the following three steps:

Step 1. Impute $y_{\text{mp,mis}}$ given $\Psi, \mu_1, \ldots, \mu_{\text{s}}, \beta, \Sigma$, and y_{obs}.

Step 2. Draw (Ψ, β, Σ) given $\mu_1, \ldots, \mu_{\text{s}}$, y_{obs}, and $y_{\text{mp,mis}}$.

Step 3. Draw $(\mu_1, \ldots, \mu_{\text{s}})$ given Ψ, β, Σ, y_{obs}, and $y_{\text{mp,mis}}$.

For the monotone Gibbs sampler for our hierarchical model, as with the single-survey MDA approach of Rubin and Schafer (1990), one need impute only enough missing data to fill in the monotone pattern y_{mp} defined in (8) and not the complete rectangular data matrix. In the Gibbs sampler context, this has the effect of analytically integrating over (rather than sampling) the other missing elements in the data matrix, which tends to yield a faster-converging algorithm (Liu, Wong, and Kong 1994).

It is straightforward to implement step 1 because, given $\Psi, \mu_1, \ldots, \mu_{\text{s}}, \beta, \Sigma$, and y_{obs}, the nonresponse components of any of the respondents in $y_{\text{mp,mis}}$ is independent of that of other respondents in $y_{\text{mp,mis}}$ and the nonresponse components of any respondent in $y_{\text{mp,mis}}$ is normally distributed. This conditional distribution is easily computed using the sweep operator.

Given $\mu_1, \ldots, \mu_{\text{s}}$, y_{obs}, and $y_{\text{mp,mis}}$, (Ψ) and (β, Σ) are independent. It is again straightforward to take a draw of (β, Σ), because the problem falls in the conventional linear

regression framework with independent normal errors, as shown in equation (6). To take a draw of Ψ from the conditional distribution, we use the following theorem, which extends the result of corollary 1 of Liu (1993).

Theorem 1. For $1 \leq k \leq p$, let $\mathbf{C}_k$ be the total sum of squares and cross-products matrix of the sample $\{(y^{(j)}_{s_i,i,k}, \ldots, y^{(j)}_{s_i,i,Q})\colon i=1,\ldots,N_j;\ j=1,\ldots,k\}$ about their corresponding population means $\mu_1, \ldots, \mu_S$. Suppose that $\mathbf{C}_k^{-1}$ has the Cholesky factorization $\mathbf{C}_k^{-1} = \mathbf{L}_k\mathbf{L}_k'$, where $\mathbf{L}_k$ is a $((Q-k+1)\times(Q-k+1))$ lower triangular matrix. Let $\mathbf{H}$ be the $(Q\times Q)$ lower triangular matrix whose kth column consists of $\mathbf{L}_k\mathbf{t}_k$ as the last $(Q-k+1)$ components, where $\mathbf{t}_k = (t_{k,k}, \ldots, t_{k,Q})'$ for $k = 1, \ldots, Q$ and $\{\mathbf{t}_k\colon k = 1, \ldots, Q\}$ satisfies the following conditions:

a. $t_{i,j}$ is independent for $1 \leq j \leq i \leq Q$.
b. $t_{i,j} \sim \mathrm{N}(0,1)$ for $1 \leq j < i \leq Q$.
c. $t_{j,j} \sim \chi_{n_1+n_2+\cdots+n_j-j+1}$ for $j = 1, \ldots, Q$.

If $n_1 > Q$, then the conditional distribution of Ψ, given $\mu_1, \ldots, \mu_S$, y_{obs}, and $y_{\mathrm{mp,mis}}$, is the same as the distribution of $(\mathbf{H}\mathbf{H}')^{-1}$.

Given Ψ, β, Σ, y_{obs}, and $y_{\mathrm{mp,mis}}$, $\mu_1, \ldots, \mu_S$ are mutually independent and normally distributed. To take a draw of μ_s for $s = 1, \ldots, S$, we use the following result.

Theorem 2. For $1 \leq k \leq Q$, let $\bar{\mathbf{y}}_{s,k}$ be the $(Q-k+1)$-dimensional sample mean of the reduced set $\{(y^{(j)}_{s_i,i,k}, \ldots, y^{(j)}_{s_i,i,Q})\colon i = 1, \ldots, n_j;\ j = 1, \ldots, k; s_i = s\}$ in survey s and the Cholesky factorization $\Psi^{-1} = \mathbf{H}\mathbf{H}'$. Then, given the monotone pattern $\{y_{\mathrm{obs}}, y_{\mathrm{mp,mis}}\}$, Ψ, β, and Σ, μ_s is conditionally normally distributed with mean

$$\eta_s = [\Sigma^{-1} + \mathbf{H}\Lambda\mathbf{H}']^{-1} \times [\Sigma^{-1}\theta_s + \mathbf{H}\Lambda_s(\mathbf{h}_1'\bar{\mathbf{y}}_{s1}, \ldots, \mathbf{h}_Q'\bar{\mathbf{y}}_{sQ})'] \quad (9)$$

and covariance matrix

$$\Phi_s = [\Sigma^{-1} + \mathbf{H}\Lambda_s\mathbf{H}']^{-1}, \quad (10)$$

where $\theta_s = \beta x_s$, $\mathbf{h}_k = (h_{k,k}, \ldots, h_{Q,k})'$, $\Lambda_s = \mathrm{diag}\,(n_{s1}, \ldots, n_{s1}+\ldots+n_{sQ})$, and n_{sk} is the number of observations of pattern k in survey s.

Following Liu (1993), we can prove Theorems 1 and 2. Letting $\bar{\mathbf{y}}_s^{(k)}$ be the $(Q-k+1)$-dimensional sample mean of the reduced set $\{(y^{(k)}_{s_i,i,k}, \ldots, y^{(k)}_{s,i,Q})\colon i = 1, \ldots, n_k; s_i = s\}$ in survey s and letting Φ_k be Ψ with elements replaced by 0 except the elements in the lower-right $((Q-k+1)\times(Q-k+1))$ submatrix, we can write η_s in (9) and Φ_s in (10) as

$$\eta_s = \left[\Sigma^{-1} + \sum_{k=1}^{Q} n_{sk}\Phi_k^-\right]^{-1}\left[\Sigma^{-1}\theta_s + \sum_{k=1}^{Q} n_{sk}\Phi_k^-\bar{\mathbf{y}}_s^{(k)}\right]$$

and

$$\Phi_s = \left[\Sigma^{-1} + \sum_{k=1}^{Q} n_{sk}\Phi_k^-\right]^{-1},$$

where Φ_k^- is Φ_k with the lower-right $((Q-k+1)\times(Q-k+1))$ submatrix replaced by its inverse.

The computer implementation is written in Fortran and C and was developed from an earlier program designed for imputation for single surveys. For the example in Section 4, the final program requires several hours to run (simulating five Gibbs sampler sequences for 100 iterations each) on a Sun Sparc workstation.

4. EXAMPLE: MISSING QUESTIONS IN PRE-ELECTION POLLS

4.1 The Problem and the Data

In a study of public opinion changes in the 1988 U.S. Presidential election campaign, Gelman and King (1993) analyzed data from 51 national opinion polls conducted by nine different major polling organizations during the 180 days preceding the election. One of the major purposes of the study was to examine changes in vote intentions (Bush, Dukakis, or undecided/other) over time for different subgroups of the population (e.g., men and women, self-declared Democrats, Republicans, and independents, low-income and high income). The changes were studied by constructing simple graphs of average vote intentions over time for different subgroups and also by tracking changes in coefficients of logistic regression models predicting vote intention in terms of variables such as sex, party identification, income, and so forth.

Performing these analyses required some care in handling the missing data, because not all questions of interest were asked in all surveys. For example, respondent's self-reported ideology (liberal, moderate, or conservative), a key variable, was missing in 10 of the 51 surveys, including our only available surveys during the Democratic nominating convention. Questions about the respondent's views of the national economy and of the perceived ideologies of Bush and Dukakis were asked in fewer than half of the surveys, and they were excluded from that analysis. Gelman and King (1993) used a mixture of available-case and complete-case methods (see Little and Rubin 1987), with available-case for the time-series plots by subgroup and complete-case for the regressions. Compared to complete-case inference, these analyses are more difficult to set up—one must examine the missing-data pattern to decide what information can be conveniently used in the analysis—and the results are more difficult to interpret, because different findings are based on different subsets of the data.

We wish to multiply impute responses for the missing questions in these and similar surveys so that the analyses for the purpose of political science need not be complicated with concerns about missing data. For example, if imputations were available, then we would not have to choose between logistic regression models that are fit to all the surveys but do not include respondent's ideology as a predictor, or include ideology but not to the surveys during the Democratic convention. Our goal in imputation is not to "get something for nothing" but rather to express the increased uncertainty due to missing data in a form that is accessible and convenient for subsequent analyses.

To this end, we fit the aforementioned multiple-imputation model to the data from the 51 pre-election surveys, using the 13 variables listed in Table 1 These included the outcome variable of interest (Presidential vote preference), the variables that were believed to have the strongest relation to vote preference, and several demographic variables that were fully observed or nearly so, which would have the effect of explanatory variables in the imputation. We also include in our analysis the date at which each survey was conducted.

There were 72,546 missing values out of 607,417 possible item responses, and an additional 249,127 missing values corresponding to questions that were not asked.

The program outputs a monotone missing-data pattern, displayed in Figure 1. Fewer than a third of the missing values in the data matrix needed to be filled in to achieve this monotone pattern.

4.2 Use of the Continuous Model for Discrete Responses

There is a natural concern when using a continuous imputation model for survey responses coded at varying levels of discretization. Some variables in our analysis (sex, ethnicity) are coded as unordered and discrete; others (vote intention, education) are ordered and discrete; and others (age, income, and the opinion questions on 1–5 and 1–7 scales) are potentially continuous but are coded as ordered and discrete. We recode the responses from different survey organizations as appropriate so that the responses from each question fall on a common scale. (For example, for the surveys in which the "perceived ideology" questions are framed as too-liberal/just-right/too-conservative, the responses are recoded based on the respondent's stated ideology.)

There are several possible ways to adapt a continuous model to impute discrete responses; from the most elaborate to the simplest, these include (1) modeling the discrete responses conditional on an underlying continuous variable (e.g., multinomial probit), (2) modeling the data as continuous and then using some approximate procedure to impute discrete values for the missing responses, and (3) modeling the data as continuous and imputing continuous values (Schafer 1997). We follow the third, simplest approach. In our example, little is lost by this simplification, because the most "discrete" variables (sex, ethnicity, vote intention) are fully observed or nearly so, whereas the variables with the most nonresponse (the opinion questions) are essentially continuous. When it is necessary to have discrete values (as for Fig. 5 in Sec. 4.7), we round off the continuous imputations, essentially using the second approach when it appears necessary.

4.3 Accounting for Survey Design and Weights

The surveys were performed by random-digit dialing (with the exception of the four Roper polls, which were in-person interviews), with one adult selected from each sampled household. The respondents for each survey were assigned weights based on sampling and poststratification (see Voss, Gelman, and King 1995 for details). These were not used in the imputation procedure, because the variables on which the weights were based were by and large already included in the imputation model. Thus the weights do not provide additional information about the missing responses, and the imputation model is proper in the sense of Rubin (1996).

We do, however, use the survey weights when computing averages, to obtain unbiased estimates of population averages unconditional on the demographic variables (i.e., weighting has the effect of poststratification on these variables). Because of the simplicity of the sampling schemes, further adjustments (e.g., for clustering) were not required. We also restricted our analysis to the respondents who stated that they were registered or were likely to vote.

If we were to include the known poststratification information (which is encoded in the weights) in the imputation analysis, then we would be able to reduce the between-survey variance in the parameters corresponding to the variables on which the weights were based. For example, two of the surveys oversample blacks, and so fitting the multiple imputation model to the data without correcting for the weights gives an estimate of proportion black varying from about 10%–33% among polls. After correcting for weights, the range reduces to 10%–16%. In our example, the variables used in the weights are all fully observed or nearly so, and so the added variability in some of the model parameters has little effect on the imputations themselves.

4.4 Presentation of Results

Before fitting the full model, we first fit the version with no survey-level variables (i.e., treating all of the surveys as exchangeable, ignoring their time order). We then demonstrate a graphical model-checking method, which can be applied routinely, that displays the failure of the exchangeable model. Section 4.5 presents results for the more appropriate model that includes time trends, Section 4.6 presents cross-validation checks for that model, and Section 4.7 compares inferences from available-case and multiple-imputation analyses. Our program displays the results of the imputation for the 51 surveys with a separate graph for each variable; we illustrate in Figure 2 with two of the variables: "income" and "perceived ideology of Dukakis." First, consider Figure 2(a), "income." Each symbol on the graph represents a different survey, plotting the estimated average value of income in the survey versus the date of the survey, with the symbol itself indicating the survey organization. The size of the symbol is proportional to the fraction of survey respondents who responded to the particular question, with the convention that when the question is not asked (indicated by circled symbols on the graph), the symbol is tiny but not of zero size. The vertical bars show ± 1 SE in the posterior mean, where the standard error is the square root of the between-imputation variance plus the average within-imputation sampling variance (as in Rubin 1987). Finally, the inner brackets on the vertical bars show the within-imputation standard deviation alone. All complete-data means and standard deviations are weighted.

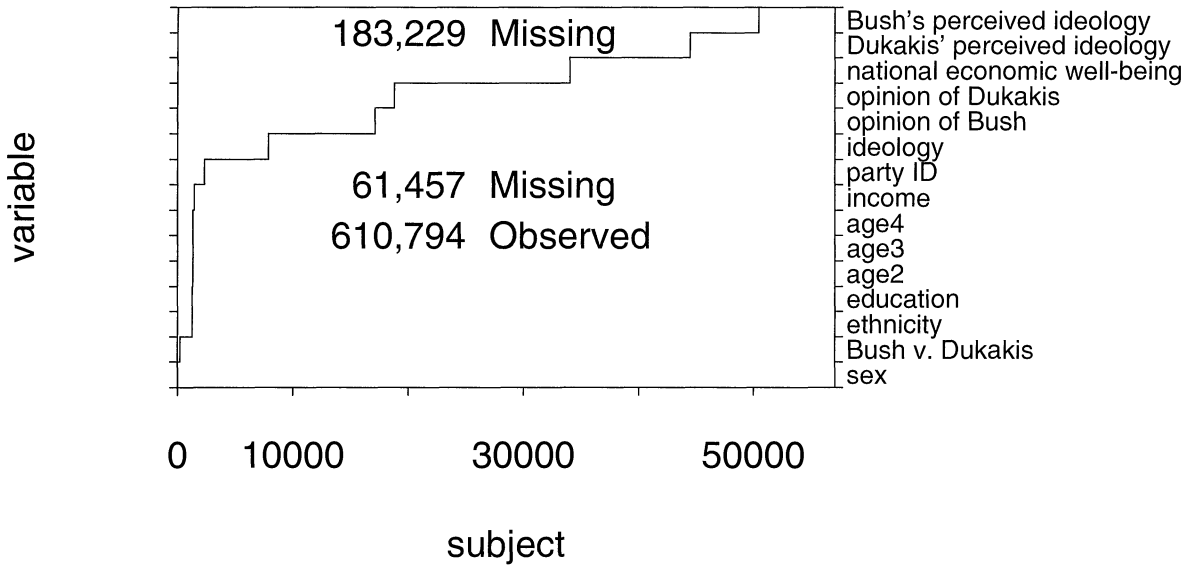


Figure 1. Monotone Data Pattern for the Pre-Election Polls, as Output by the Multiple-Imputation Program.

For surveys in which the question was asked, the within-imputation variance almost equals the total variance, which makes sense, because when a question was asked, most respondents answered (see Table 1). The multiple-imputation procedure makes very weak statements about missing income responses, which makes sense, because income is not highly correlated with the other questions. However, even the surveys in which this question was asked have nonzero standard errors, because of the finite sample sizes of the surveys.

Figure 2(a) also shows some between-survey variability in average income, from 31K to 37K—more than can be explained by sampling variability, as is indicated by the error bars on the surveys for which the question was asked. Because we do not believe that the average income among the population of registered or likely voters is changing that much, the explanation must lie in the surveys. In fact, different survey organizations use different codings for incomes (e.g., 0–10K, 10–20K, 20–30K, etc., or 0–7.5K, 7.5–15K, 15–25K, etc.). Because the point of our method is to produce imputations close to what the surveys would look like if all the questions had been asked and answered, rather than to adjust all the observed and unobserved data to estimate population quantities, this variability is reasonable. The large error bars for average income for the surveys in which the question was not asked reflect the large between-survey variation in average income, which is captured by our hierarchical model. For this study, we are interested in income as a predictor variable rather than for its own sake, and we are willing to accept this level of uncertainty.

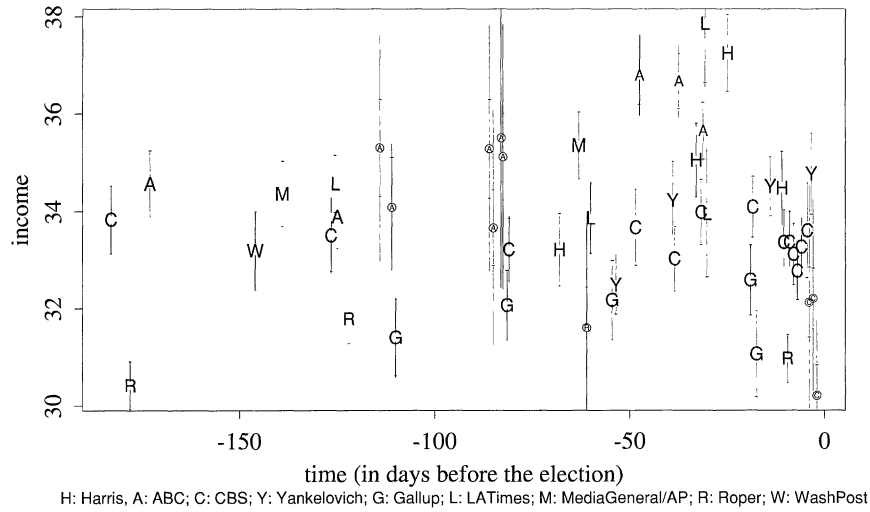
4.5 Model Checking and Improvement

Figure 2(b) shows a similar plot for Dukakis's perceived ideology. This graph shows a serious flaw in the model:

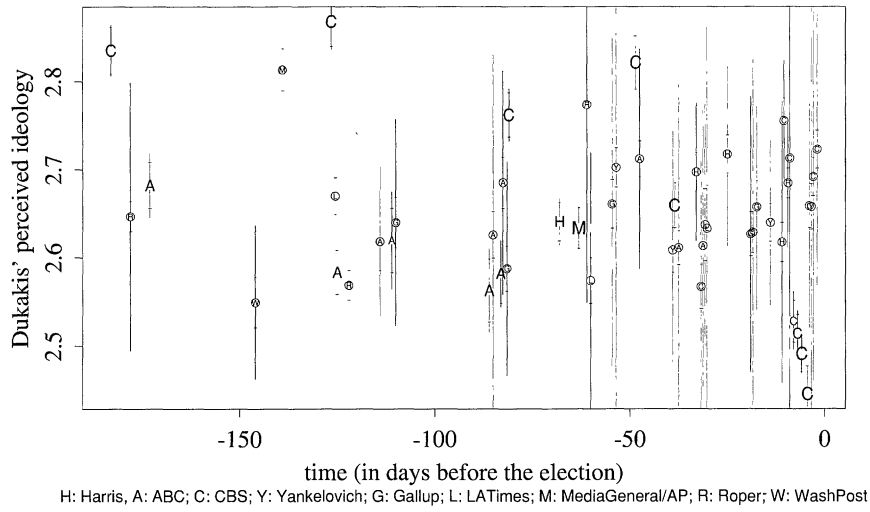
Table 1. Survey Questions Used in the Multiple Imputation Study

Question	*Range of responses*	*No. of surveys*	*Rate of item nonresponse*
Vote intention	1 (Bush), 1.5 (undecided), 2 (Dukakis)	48	.15
Sex	1 (male), 2 (female)	51	0
Age	18–65+ years	49	.08
Education	1 (no high school)–5 (graduate school)	45	.03
Ethnicity	1 (white), 1.5 (other), 2 (black)	51	.03
Income	0–100+ thousands of dollars	41	.12
Party identification	1 (strong Republican)–7 (strong Democrat)	50	.07
Ideology	1 (very liberal)–5 (very conservative)	41	.10
Opinion of Bush	1 (very favorable)–5 (very unfavorable)	36	.30
Opinion of Dukakis	1 (very favorable)–5 (very unfavorable)	36	.30
View of economy	1 (very good)–7 (very bad)	20	.14
Perceived ideology: Bush	1 (very liberal)–5 (very conservative)	10	.30
Perceived ideology: Dukakis	1 (very liberal)–5 (very conservative)	16	.38

NOTE: "Number of surveys" is the number of surveys (out of 51) in which the question was asked, and "Rate of item nonresponse" is for that question among those surveys in which it was asked. Demographic questions such as sex, ethnicity, education, and income, which are nearly fully observed, are essentially used as explanatory variables in the imputation. All variables were coded as above, except for age, which was discretized into categories 18–29, 30–45, 46–54, and 65+ (i.e., the continuous "age" variable was replaced by indicator variables for three of the four age categories); and income, which was treated as a continuous variable with values assigned from the approximate median for each response category (e.g., 0–10K set to 7K, 10–20K set to 15K, . . . , 60K+ set to 80K, 100K+ set to 125K). Also, when the perceived ideology of Bush or Dukakis was stated to be "too liberal," "too conservative," or "just right," the response for the perceived ideology of the candidate was imputed to 2, 4, or the respondent's answer to the "ideology" question.



(a)



(b)

Figure 2. Estimates and ±1 SE Bars for the Population Mean Response for Two Questions—(a) Income and (b) Perceived Ideology of Dukakis, Over Time—for the Model That Does Not Include Time as a Covariate. Each symbol represents a different survey, with different letters indicating different survey organization. The size of the letter indicates the number of responses to the question, with large-sized letters for surveys with nearly complete response and small-sized letters for surveys with few responses. Circled letters indicate surveys for which the question was not asked; note that the estimates for these surveys have much larger standard errors. The inner brackets on the vertical bars show the within-imputation standard deviation for the estimated mean from each survey. Note the anomaly in (b), which indicates a model error: the surveys with responses (the large letters) show a trend over time, whereas the surveys without responses (the small, circled letters) do not. This problem was fixed by including time as a covariate (see Fig. 3).

The surveys in which the question was asked (indicated by uncircled letters) show a strong trend downward over time (toward a perceived ideology of "liberal"), whereas the surveys in which the question was not asked (indicated by circled letters) are approximately constant over time. This indicates that the trend in Dukakis's perceived ideology is not captured by the regression model, which ignores time, from the other survey responses. Plots for other survey responses (most notably, "opinion of Dukakis") show similar time trends not captured by the multiple imputations under the model that does not include time as a predictor.

The obvious solution to this problem is to put time trends into the model, which we do by including time as a survey-level covariate—that is, a known variable x in model (5). Figure 3 shows the plots corresponding to Figure 2 under the old model; there are no apparent problems now. If one were further interested in exploring survey effects, then one could add indicator variables for survey organizations as additional covariates x. This would require a further level of hierarchical modeling, however, as not all questions are asked by all survey organizations. Thus some pooling or partial pooling would be required for the coefficients of x in the multivariate regression.

In general, further graphical checks on the model fit, such as residual plots, would be appropriate, depending on the purpose to which the model would be used. (That is, it might

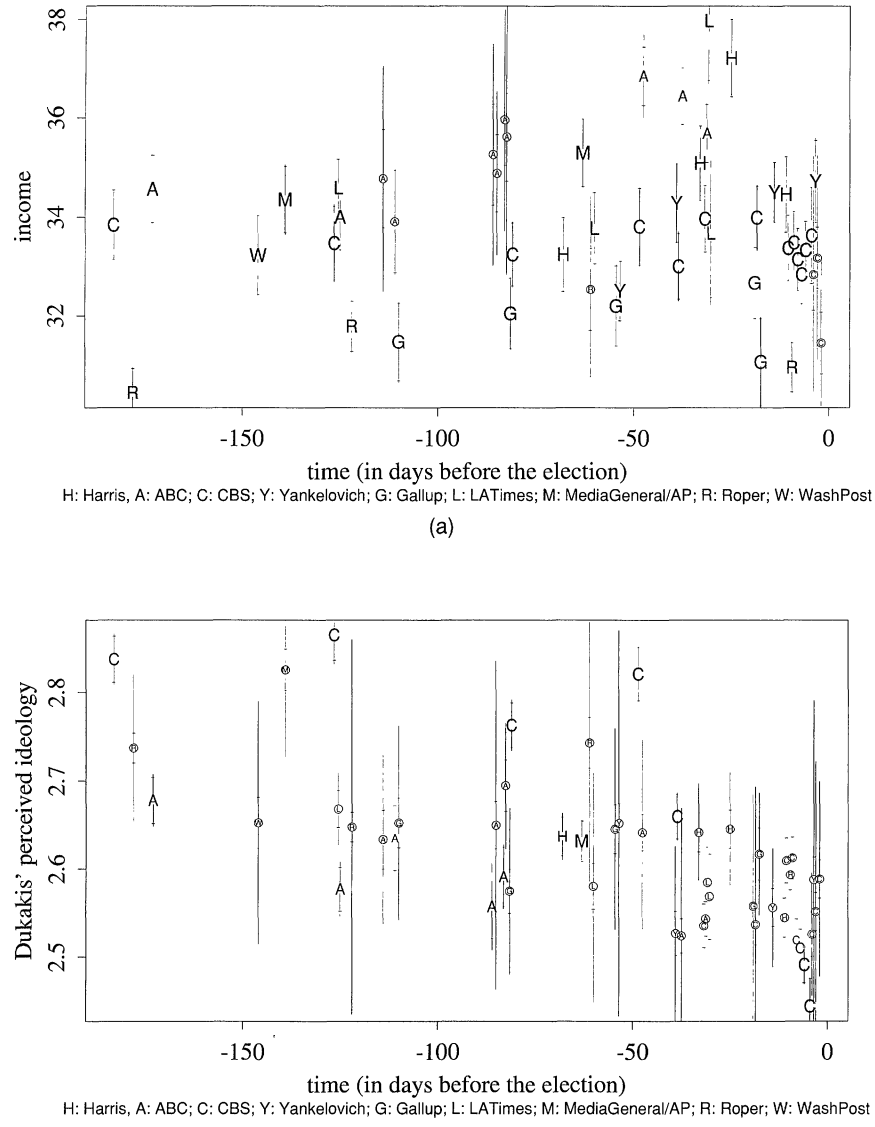


Figure 3. Estimates and ±1 SE Bars for the Population Mean Response for Two Questions—(a) Income and (b) Perceived Ideology of Dukakis, Over Time—for the Model That Includes Time as a Covariate. Each symbol represents a different survey, with different letters indicating different survey organization. The size of the letter indicates the number of responses to the question, with large-sized letters for surveys with nearly complete response and small-sized letters for surveys with few responses. Circled letters indicate surveys for which the question was not asked; note that the estimates for these surveys have much larger standard errors. The inner brackets on the vertical bars show the within-imputation standard deviation for the estimated mean from each survey. Note that the anomaly in Figure 2(b) has been corrected.

be appropriate for these tests to be performed by the users as well as the creators of the multiple imputations.) In any particular application, we imagine that in-depth examination of the imputed data would be useful for discovering ways to improve the imputation model.

4.6 Cross-Validation Checks

4.6.1 Ignorable Nonresponse. To test the model in another way, we created a new dataset by removing the "party identification" question from half of the surveys in which it was asked, and then removing the responses for that question from a random selection of a third of the individuals in the remaining surveys. The resulting nonresponse pattern for "party identification" is then comparable to the items on the bottom of Table 1. This nonresponse mechanism is ignorable, in the sense that the nonresponse *pattern* provides no information about the missing data *values*. We then ran the multiple-imputation program on this new dataset and compared the imputed values of party identification to the true values that were artificially deleted. This is a serious check of our method, because party identification is the best-known predictor of vote intention and is highly correlated with many of the other questions.

We checked the multiple imputations by comparing them to the withheld data (the responses to the party identification question that were withheld from the analysis) in two ways: averaged over surveys and as individual responses. Figure 4(a) displays, for each poll, the average response to the party identification question (on the y-axis) versus the average from the multiply imputed datasets. Open circles indicate

surveys where the party identification question was (artificially) completely missing; in solid circles, the question was missing from about $\frac{1}{3}$ of the units. To indicate uncertainty in the forecasts, horizontal error bars display ± 1 SD from the between-imputation variability. Predictions are of course much more accurate for the surveys in which responses to the question were available. The imputations have approximately zero mean error; that is, $\mathrm{E}(\overline{y}^{\mathrm{actual}}|\overline{y}^{\mathrm{pred}}) \approx \overline{y}^{\mathrm{pred}}$, where $\overline{y}^{\mathrm{pred}}$ is the mean of the five imputed values of $\overline{y}$. In addition, the uncertainties in the imputed means are reasonably calibrated: about $\frac{2}{3}$ of the true values fall within one predicted standard error.

We also check the predictions of individual responses to the party identification question. For each individual for whom we artificially removed the response to that question, we compare the actual response to the five multiply imputed responses. If the data had been simulated from the model, then we would expect the actual responses and the multiple imputations to have the same distribution, so that if one ranked the actual response along with five random imputations, then all six possible orderings (actual response lowest, second lowest, ..., highest) would be equally likely. The first three columns of Table 2 present the actual cross-validation results for the party identification question, separating the surveys with (artificially created) complete nonresponse and $\frac{1}{3}$ nonresponse. In both cases, the six rankings are quite close to equally likely, meaning that the spread of the five imputations is approximately calibrated to the actual predictive uncertainty. In comparison, if predictions were systematically overconfident, then the extreme categories would have higher frequencies; if predictions were systematically underconfident, then the extreme categories would have lower frequencies. If, in addition, the imputations were systematically too high or too low, then the frequencies in the six categories would show a decreasing or increasing trend, respectively.

4.6.2 Nonignorable Nonresponse. We repeat the foregoing cross-validation simulation with a highly *nonignorable* nonresponse pattern: as before, we remove the party identification question from half of the surveys, but in the remaining surveys, we remove the response from 10% of (self-declared) Republicans, 30% of Independents, and 50% of Democrats. This once again yields an approximate $\frac{1}{3}$ item nonresponse rate, but with nonresponse probabilities that depend on the now-unobserved responses. We then repeat the aforementioned calibration checks; because we have disproportionately removed the responses from Democrats, we expect the imputed values to be too low, compared to the true responses. (Party identification is coded from 1 for strong Republican to 7 for strong Democrat.)

Figure 4(b) displays, for each poll, the average response to the party identification question (on the y-axis) versus the average from the multiply imputed datasets. The imputations are clearly too low both for the surveys in which item nonresponse was created (solid circles) and those from which all responses to the question were removed (open circles). Interestingly, the standard errors are much too small for the solid circles but are closer to calibrated (although still too small) for the open circles. This suggests that the hierarchical between-survey variability of the model protects the imputations from being very overconfident in surveys

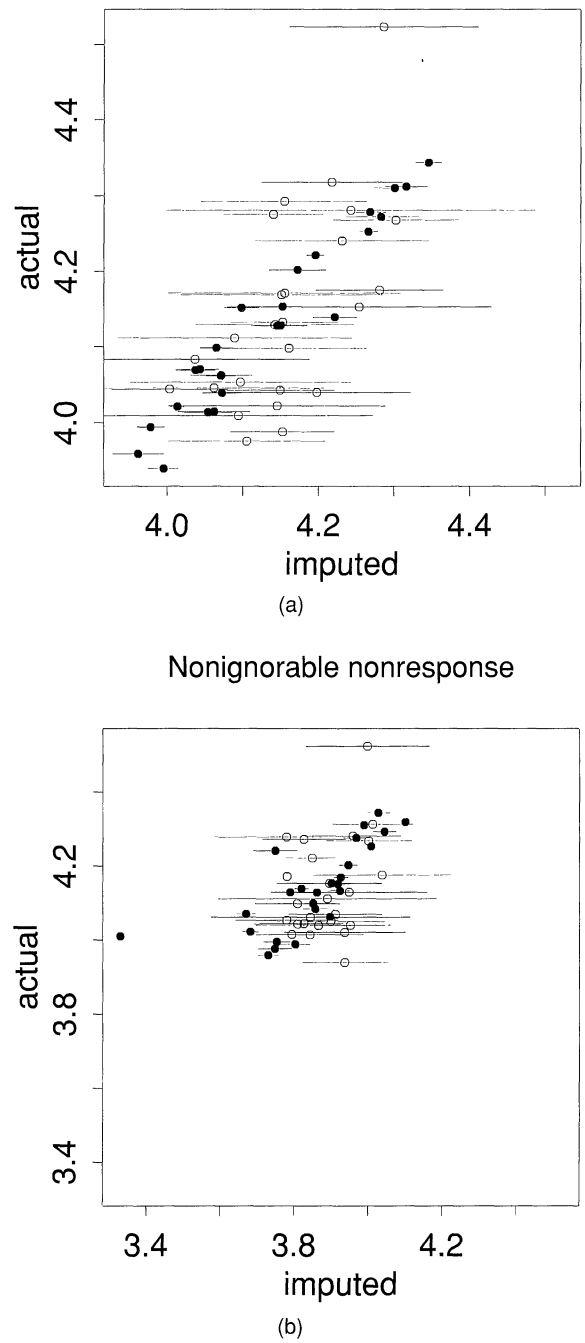


Figure 4. Cross-Validation Checks for the Party Identification Question With (a) Ignorable and (b) Nonignorable Nonresponse. The question was artificially removed from $\frac{1}{2}$ of the surveys and $\frac{1}{3}$ of the responses from the remaining surveys, and the missing values were multiply imputed by our computer program. Two different cross-validations were performed, one with ignorable nonresponse and one with nonignorable nonresponse. For each, the predicted average response for each survey (from the average of the multiple imputations) is compared to the true value, with horizontal error bars indicating ± 1 standard deviation of between-imputation variability. Open circles indicate surveys in which all the responses to the question were deleted; solid circles indicate surveys in which a random selection of $\frac{1}{3}$ of the responses were deleted. The diagonal line corresponds to equality between actual and imputed values. Note that the ranges on the horizontal and vertical axes are the same on each graph but differ between graphs.

Table 2. Cross-Validation of Individual Responses for the Party Identification Question

Rank of true response among the five imputations	Ignorable nonresponse		Nonignorable nonresponse	
	Proportion of cases (for polls with complete nonresponse)	Proportion of cases (for polls with $\frac{1}{3}$ nonresponse rate)	Proportion of cases (for polls with complete nonresponse)	Proportion of cases (for polls with $\frac{1}{3}$ nonresponse rate)
0	16.5%	16.8%	12.3%	5.9%
1	17.9%	17.7%	16.2%	8.6%
2	16.3%	15.8%	16.3%	11.1%
3	17.1%	16.8%	16.6%	16.4%
4	16.8%	16.7%	18.8%	23.5%
5	15.3%	16.2%	19.8%	34.4%

NOTE: Columns 2 and 3 correspond to a simulation with *ignorable nonresponse*; columns 4 and 5 correspond to a simulation with *nonignorable nonresponse*. The responses for this question were removed entirely from $\frac{1}{2}$ of the surveys (chosen at random) and from $\frac{1}{3}$ of the individuals in the other surveys. For each of the removed responses, we recorded its ranking among the five imputations (0 means the true value is lower than all five imputations, 1 means it is lower than four of the five imputations, $\cdots$, and 5 means it is higher than all five imputations), breaking ties randomly. The table records the percentage of values in each category, considering separately the polls with complete nonresponse and partial nonresponse. For each simulation, if the imputation model were correct, then we would expect about 16.7% in all categories.

for which the question was not asked, even if the nonresponse is nonignorable and the imputations are, on average, quite biased.

The last two columns of Table 2 present the cross-validation results for the predictions of individual responses to the missing party identification question. Once again, we separate the polls with complete nonresponse from those with item nonresponse. As expected, the imputations are not calibrated: for example, in the polls with item nonresponse, 34.4% of the true values of the "missing" party identification are higher than all five multiply imputed values. Under the model, we would expect only 16.7%. The direction of this bias is as expected, given that a disproportionate number of high values (Democrats) were removed, and this was not accounted for in the model. Once again, however, the lack of calibration is not nearly as bad for the polls in which all the questions were removed; here, only 19.8% of the true values were greater than all five multiple imputations. These results suggest that the aspect of our imputation model that is the most vulnerable to nonignorable nonresponse is the traditional within-survey imputation, not the new hierarchical model for between-survey variation.

4.7 Comparison of Available-Case and Multiple Imputation Analyses

We conclude by replicating one of the analyses of Gelman and King (1993): the plots of political preference by subgroup, over time. Each of the plots in Figure 5 displays the estimated changes in estimated support for Bush over time for different groups of the population, as characterized by survey responses. The population is separated in turn according to political party identification (Republican, Democrat, and Independent/other/no-answer), ideology (conservative, liberal, and moderate/no-answer), income (under \$20,000, \$20,000–\$50,000, and over \$50,000), and view of the economy (positive, intermediate or negative). We include the two political variables because they are the most strongly predictive of vote preference, we include income because it has a relatively high rate of nonresponse for a demographic variable, and we include view of the economy as an important variable that was asked in fewer than half of the surveys.

The plots on the left column of Figure 5 display the results based on an available-case analysis, using, for each plot, only the surveys in which the corresponding question was asked and only the individuals who responded to those items. For each poll, error bars show ± 1 SE, estimated from the weighted mean of the respondents. The plots in the right column of Figure 5 display the corresponding results using the multiply imputed datasets, with standard errors including both within- and between-imputation variation.

How do the available-case and multiple-imputation analyses differ? The most striking pattern is during the Republican convention (about 115 days before the election), when the available polls do not ask the "ideology" or "income" questions. The available-case analyses must skip this point, whereas the analyses from the imputations show the different subgroups to be moving together over time. This behavior revealed by the analysis of the multiply imputed data makes sense politically. The fact that public opinion shifts are generally uniform across the population is documented elsewhere (Gelman and King 1993). Page and Shapiro (1992) used the term "parallel publics" for this behavior and discussed it extensively in many aspects of U.S. public opinion.

5. CONCLUSION

The method of multiple imputation, analysis, and diagnostics based on a hierarchical regression model achieves the goal of generalizing available algorithms for single-survey imputation to attack the problem of imputation for several surveys or for several strata or clusters within a single survey. We perform the computations using an iterative algorithm (the Gibbs sampler) that alternately performs imputation at the single-survey level and estimates parameters using information available from all the surveys. The results have the Bayesian property of compromising between the approaches of no pooling and complete pooling of surveys. The estimated between-survey variation is part of the multiple-imputation variation, which typically yields

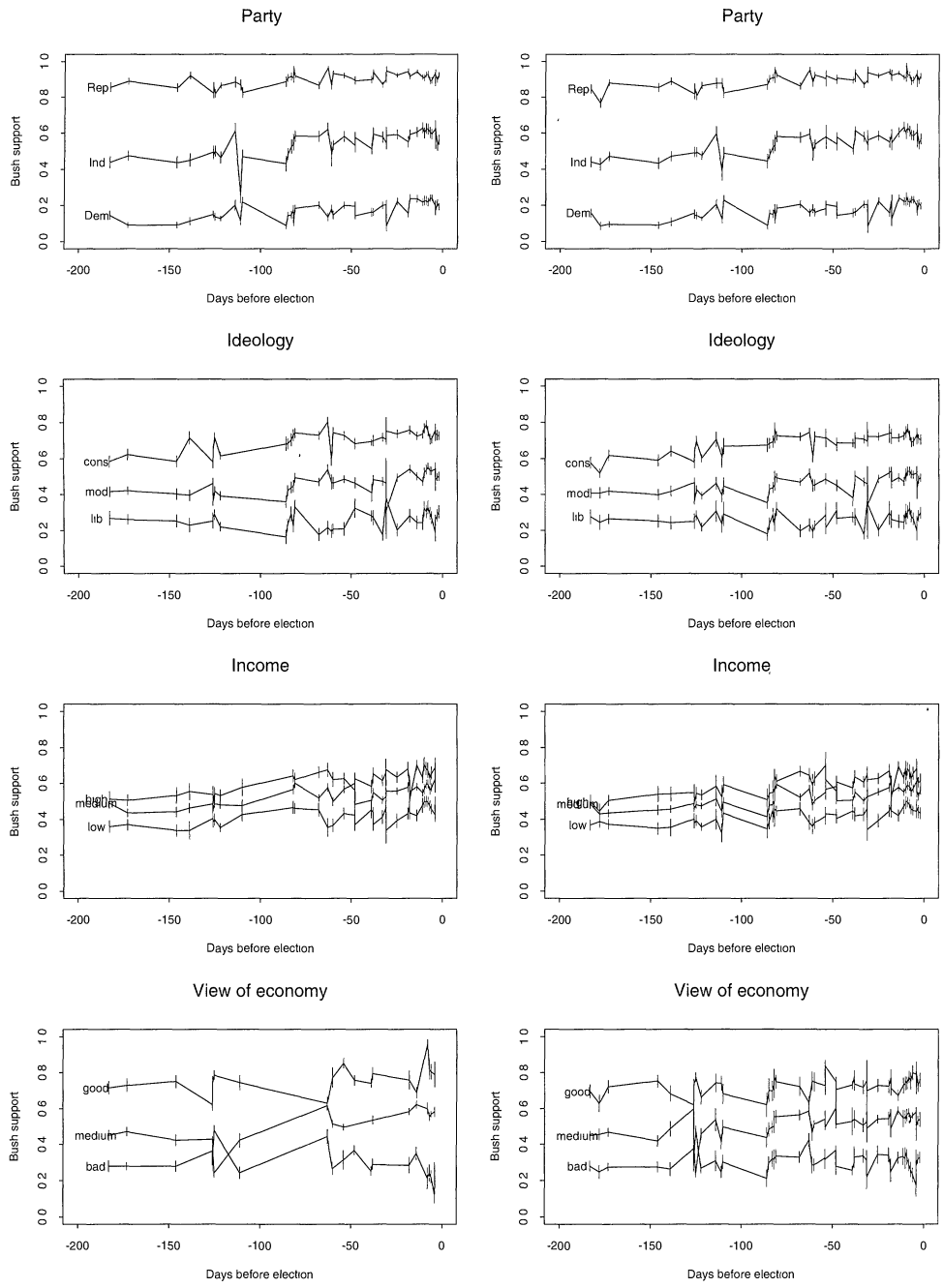


Figure 5. Comparison of Available-Case and Multiple-Imputation Analyses. The plots on the left side display available-case analyses; those on the right side are the corresponding analyses based on the multiple imputations. The most notable differences appear during the Republican convention (about 115 days before the election), when the available polls did not include ideology and income questions.

wide posterior intervals for questions that were not asked in a particular survey. Information from survey weights is incorporated by including in the analysis the variables on which the weights were based, and then reweighting individual responses (observed and imputed) to estimate population quantities.

Cross-validation studies show that the ignorable model performs well for ignorable nonresponse but poorly under strongly nonignorable nonresponse. The most immediate application of these methods is for problems like our election study—an analysis of a series of independent cross-sectional surveys in which not all questions are asked in all surveys, and with relatively low rates of item nonresponse for the questions of primary interest. Note also that the ability to include more variables in the imputation model (by including variables that are not asked in all the surveys) should give our model more flexibility to handle item nonresponse. (See Rubin 1996, Sec. 2.6, for a discussion of why the missing at random assumption is in general more reasonable if more variables are included in the model.)

Once imputations have been obtained, the completed datasets can be analyzed using complete-data methods of inference. Before doing so, however, it is advisable to summarize the results of the imputations graphically, using symbol

sizes to indicate the fraction of missing data in the different surveys. Graphs of posterior inference for mean responses, plotted against time, survey organization, and other survey-level variables, are crucial for identifying variables that should be included in the hierarchical model.

When performing imputations, questions always arise about the adequacy of the model used to create the imputations. Some aspects of model adequacy can be addressed internally, as discussed earlier, but the ultimate test is to compare the results of analyses of substantive interest to what would be obtained using various methods of imputation- or nonimputation-based analysis. This is what was done in the example of the pre-election polls. Ultimately, to choose any data analysis procedure is to make a decision, and some of the purposes of multiple imputation are to make the assumptions behind that decision more transparent, account for as much uncertainty as possible, and reduce the complexities of subsequent substantive analyses.

[Received June 1997. Revised January 1998.]

REFERENCES

Belin, T. R., Diffendal, G. J., Mack, S., Rubin, D. B., Schafer, J. L., and Zaslavsky, A. M. (1993), "Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation" (with discussion), *Journal of the American Statistical Association*, 88, 1149–1166.

Dominici, F., Parmigiani, G., Reckhow, K. H., and Wolpert, R. L. (1996), "Combining Information From Related Regressions," unpublished manuscript submitted to *Journal of Agricultural, Biological, and Environmental Statistics*.

DuMouchel, W. M., and Harris, J. E. (1983), "Bayes Methods for Combining the Results of Cancer Studies in Humans and Other Species" (with discussion), *Journal of the American Statistical Association*, 78, 293–315.

Efron, B., and Morris, C. (1975), "Data Analysis Using Stein's Estimator and Its Generalizations," *Journal of the American Statistical Association*, 70, 311–319.

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

Franklin, C. H. (1989), "Estimation Across Data Sets: Two-Stage Auxiliary Instrumental Variables Estimation (2SAIV)," *Political Analysis*, 1, 1–24.

Gatsonis, C. A, Normand, S. L., Morris, C., and Liu, C. H. (1992), "Geographic Variation of Procedure Utilization: A Hierarchical Model Approach," *Medical Care*, 31, YS54–59.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, New York: Chapman and Hall.

Gelman, A., and King, G. (1993), "Why Are American Presidential Election Campaign Polls So Variable When Votes Are So Predictable?" *British Journal of Political Science*, 23, 409–451.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Kish, L. (1965), *Survey Sampling*, New York: Wiley.

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

Liu, C. (1993), "Bartlett's Decomposition of the Posterior Distribution of the Covariance for Normal Monotone Ignorable Missing Data," *Journal of Multivariate Analysis*, 46, 198–206.

——— (1995), "Missing Data Imputation Using the Multivariate t Distribution," *Journal of Multivariate Analysis*, 53, 139–158.

——— (1996), "Bayesian Robust Multivariate Linear Regression With Incomplete Data," *Journal of the American Statistical Association*, 91, 1219–1227.

Liu, C., and Rubin, D. B. (1998), "Ellipsoidally Symmetric Extensions of the General Location Model for Mixed Categorical and Continuous Data," unpublished manuscript submitted to *Biometrika*.

Liu, J., Wong, W. H., and Kong, A. (1994), "Covariance Structure of the Gibbs Sampler With Applications to the Comparisons of Estimators and Augmentation Schemes," *Biometrika*, 81, 27–40.

Meng, X. L. (1994), "Multiple-Imputation Inferences With Uncongenial Sources of Input" (with discussion), *Statistical Science*, 9, 538–573.

Page, B. I., and Shapiro, R. Y. (1992), *The Rational Public*, Chicago: University of Chicago Press.

Raghunathan, T. E., and Grizzle, J. E. (1995), "A Split Questionnaire Survey Design," *Journal of the American Statistical Association*, 90, 54–63.

Rao, J. N. K. (1996), "On Variance Estimation With Imputed Survey Data," *Journal of the American Statistical Association*, 91, 499–505.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

——— (1980), "Using Empirical Bayes Techniques in the Law School Validity Studies" (with discussion), *Journal of the American Statistical Association*, 75, 801–827.

——— (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

——— (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91, 473–489.

Rubin, D. B., and Schafer, J. L. (1990), "Efficiently Creating Multiple Imputations for Incomplete Multivariate Normal Data," in *Proceedings of the Statistical Computing Section of the American Statistical Association*, pp. 83–88.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, New York: Chapman and Hall.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 528–550.

Voss, D. S., Gelman, A., and King, G. (1995), "Pre-Election Survey Methodology: Details From Nine Polling Organizations, 1988 and 1992," *Public Opinion Quarterly*, 59, 98–132.