Models for Nonresponse in Sample Surveys

RODERICK J. A. LITTLE*

The literature on the analysis of incomplete data using models is reviewed in the context of nonresponse in sample surveys. The modeling approach provides a large body of methods for handling unit and item nonresponse, some of which cannot be derived from the randomization theory of inference for surveys. Key concepts from the literature on incomplete data, such as factorizations of the likelihood for special data patterns, the EM algorithm for general data patterns, and ignorability of the response mechanism, are discussed within the survey context. Model-based procedures are related to common methods for handling nonresponse in surveys, such as weighting or imputation of means within subclasses of the population.

KEY WORDS: Superpopulation models; Maximum likelihood; Bayes theorem; Missing data.

1. INTRODUCTION

The literature on inference from sample surveys is complex and distinguishes many subtle differences in philosophy. For a recent discussion within the context of survey nonresponse, see Brewer and Särndal (1979). In the discussion of that article, Little and Rubin (1979) emphasize a single distinction as being of crucial importance, namely that between what might be termed the *randomization* and the *model-based* approaches. In the randomization approach, the population values are treated as fixed, and inferences are based on the probability distribution used to select the sample. In the modeling approach, the population values are treated as realizations of random variables that are distributed according to some model. The model distribution forms the basis for inferences, and the sample selection procedure has an ancillary role, namely to avoid selection bias. The two approaches can be illustrated by the following familiar example:

Example 1.1 Simple random sampling with a single completely observed variable. Let N denote the population size, x_i the value of a variable x for unit i, $i = 1, \ldots, N$, and let

$$\bar{X} = \sum_i^N x_i/N$$

denote the population mean of x. Define the sample indicator function δ, where $\delta_i = 1$ if unit i is selected, $\delta_i = 0$, otherwise. Finally, let $n = \sum_i^N \delta_i$ be the sample size. *Randomization* inference is based on the sample mean

$$\bar{x} = \sum_{i=1}^N \delta_i/n,$$

considered as a random function of the sample indicators $\delta = (\delta_1, \ldots, \delta_N)$, with the population values $x_1, \ldots, x_N$ fixed. For moderately large samples, $\bar{x} = \bar{x}(\delta)$ is normal with mean $\bar{X}$ and variance $(1 - n/N)S^2/N$, where S^2 is the population variance of the x_i^2's. In symbols,

$$\bar{x}(\delta) \mid (x_1, \ldots, x_N) \sim G(\bar{X}, (1 - n/N)S^2/n), \quad (1.1)$$

where G denotes the normal (Gaussian) distribution. Estimating S^2 by the sample variance s^2, (1.1) leads to confidence intervals for $\bar{X}$ in the usual way.

Equivalent *model-based* inferences for this problem treat the values x_i as independent identically distributed (iid) normal random variables with mean μ and variance σ^2; that is,

$$x_i \sim \text{iid } G(\mu, \sigma^2), \quad i = 1, \ldots, N, \quad (1.2)$$

where μ and σ^2 are *superpopulation parameters*, in contrast to $\bar{X}$ and S^2, which are *population quantities*. Two variants of the argument now arise. In the approach of Royall (1970), inferences are based on properties of estimates of $\bar{X}$ (such as $\bar{x}$) in repeated sampling from the superpopulation distribution (1.2). In the Bayesian approach of Ericson (1969), prior distributions are specified for μ and σ^2, and inferences are based on the posterior distribution of $\bar{X}$ given the data. In the latter approach, assuming moderately large samples and vague priors on μ and σ^2, we obtain by standard Bayesian theory

$$\mu \mid \text{data} \sim G(\bar{x}, s^2/n) \quad (1.3)$$

for the posterior distribution of μ, and

$$\bar{X} \mid \text{data} \sim G(\bar{x}, (1 - n/N)s^2/n), \quad (1.4)$$

for the posterior distribution of $\bar{X}$. The distribution (1.4) leads to posterior probability intervals for $\bar{X}$ that are nu-

[*] Roderick J. A. Little is Vice President, Datametrics Research Inc., 4805 Grantham Avenue, Chevy Chase, MD 20015, and Adjunct Associate Professor; The George Washington University, Washington, D.C. 20052. An earlier draft of this article was presented as an invited paper for the 13th European Meeting of Statisticians, University of Sussex, September 1980. The article was written while the author was Scientific Associate for the World Fertility Survey, International Statistical Institute, and is based in part on chapters written for the National Academy of the Sciences Panel on Incomplete Data (Little 1980). The author acknowledges useful discussions with members of the panel, and in particular with Arthur Dempster, William Madow, Richard Platek, Donald Rubin, and Frederick Scheuren. The helpful comments of an associate editor and three referees are also acknowledged.

© Journal of the American Statistical Association
June 1982, Volume 77, Number 378
Applications Section

merically identical to the confidence intervals obtained from (1.1). Royall's approach also yields similar results, with variance replaced by mean squared error in estimates of precision (Royall 1970).

For the simple random sampling design of this example, randomization theory and the normal model lead to similar results. In more realistic examples, knowledge of the population allows more efficient sampling designs, which may involve stratification, clustering, or variable probability sampling. For such designs, the difference between the randomization and modeling approaches is more marked.

The revival of interest in the modeling approach is comparatively recent and hampered by misunderstandings about the role of randomization (see Sec. 2.2). Appropriate models for surveys that take into account the sample design and are protected against specification error remain relatively undeveloped, with some notable exceptions (e.g., Scott and Smith 1969; Royall and Herson 1973; Fay and Herriott 1979). However, once a model has been specified, the modeling approach avoids many theoretical problems in the randomization approach, such as the choice of estimator or the appropriate conditioning (Holt and Smith 1979).

The great attraction of the randomization theory is that the need to specify a model is avoided. The probability distribution is known, whereas a model involves a subjective element. Unfortunately, this objectivity is lost when deviations from probability sampling occur, such as errors of coverage, response errors, or unit or item nonresponse. In the case of nonresponse, the pure randomization approach can only be saved by assuming that within recognizable subclasses of the population, nonresponse is itself a form of probability sampling (see Sec. 3.4).

The aim of this article is to describe the treatment of survey nonresponse within the framework of superpopulation modeling (Little 1980; Herzog and Rubin 1980; Rubin 1980). A considerable relevant literature exists on the treatment of missing values in statistical modeling in general. In particular, there is a large literature on maximum likelihood (ML) estimation for incomplete data. Review articles include Hartley and Hocking (1971), Orchard and Woodbury (1972), and Dempster, Laird, and Rubin (1977).

The emphasis in this literature is on the estimation of the superpopulation parameters of the model (the μ and σ^2 in our simple example). However, the methods are often easily adapted to inferences about finite population parameters, such as $\bar{X}$. In many problems, the step from inferences about superpopulation parameters (cf. expression (1.3)) to inferences about population quantities (cf. expression (1.4)) simply involves the inclusion of finite population corrections like the factor $(1 - n/N)$ in the variance of (1.4). If the sampling fraction is small, then inferences about the superpopulation parameter (for example μ) and the corresponding population quantity ($\bar{X}$) are often practically identical.

2. SUPERPOPULATION MODELS FOR SAMPLE SURVEYS WITH NONRESPONSE

2.1 Notation

In this section, Roman letters will be used to denote random variables, which are scalar, vector, or matrix valued according to context. Probability density functions of continuous random variables and distribution functions of discrete random variables will be denoted by the symbol f, and densities for various random variables will be distinguished by their arguments. Parameters will be denoted by Greek letters. For example, if x and y have a joint distribution indexed by the parameter θ, then the probability density function of the joint distribution function is denoted by $f(x, y; \theta)$. The probability density functions of the marginal distributions of x and y are denoted by $f(x; \theta)$ and $f(y; \theta)$, respectively. The probability density function of the conditional distribution of x given y is denoted by $f(x \mid y; \theta)$.

Likelihoods will be denoted by the symbol L and log-likelihoods by the symbol l. To emphasize that these are functions of the parameters with the data fixed, parameters are written as the first argument. For example, the likelihood of θ based on data (x, y) is

$$L(\theta; x, y) = \text{const.} \times f(x, y; \theta), \quad \text{for all } \theta,$$

and $l(\theta; x, y) = \ln L\ (\theta; x, y)$. Finally, the likelihood derived from the conditional distribution of x given y is written $L(\theta; x \mid y)$.

2.2 Superpopulation Models in the Absence of Nonresponse

Data from survey without nonresponse can be represented as in Figure 1, with rows representing units, and columns, variables. Shaded areas represent data. Three sets of variables can be distinguished. The *sample-design* variables represent variables assumed known for all units of the population. The *sample-indicator variable* δ is a dichotomy, such that $\delta_i = 1$ if unit i is sampled, and $\delta_i = 0$ if unit i is not sampled. The *item variables* x are recorded only for the sampled items. The set of sampled values of x is written x_s and the set of nonsampled values is written $x_{\bar{s}}$.

A superpopulation model treats x and δ as realizations of random variables with the following distributions.

1. Conditional on the sample-design variables z, the population items x have a distribution $f(x \mid z; \theta)$ indexed by a set of parameters θ.
2. Conditional on z and x, the sample-indicator function has a distribution $f(\delta \mid z, x; \phi)$ indexed by a set of parameters ϕ.

These two components combine to form the joint distribution of x and δ given z, θ, and ϕ:

$$f(x, \delta \mid z; \theta, \phi) = f(x \mid z; \theta) f(\delta \mid z, x; \phi). \quad (2.1)$$

The distribution of the data x_s and δ is obtained by in-

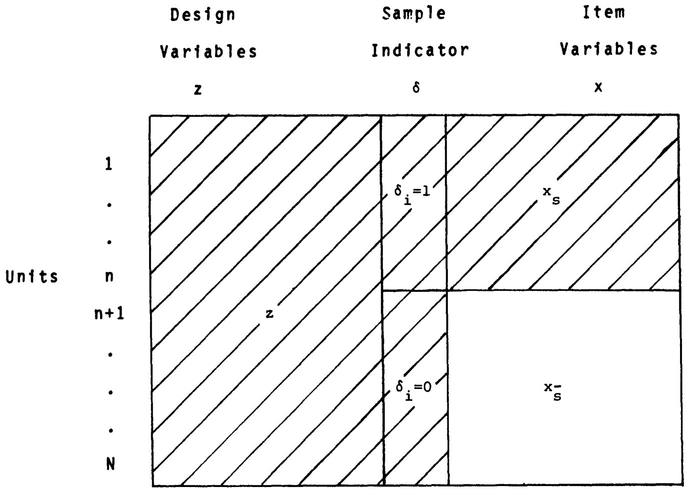

Figure 1. The Data From A Sample Survey Without Nonresponse

tegrating (2.1) over the nonsampled items $x_{\bar{s}}$:

$$f(x_s, \delta \mid z; \theta, \phi) = \int f(x_s, x_{\bar{s}} \mid z; \theta) \times f(\delta \mid z, x_s, x_{\bar{s}}; \phi) dx_{\bar{s}}. \quad (2.2)$$

A characteristic of this formulation is that the sampling distribution of δ is conditional on the population items x and is included in the model. Usually, superpopulation-model inferences are based on the marginal distribution of the sampled items:

$$f(x_s \mid z; \theta) = \int f(x_s, x_{\bar{s}} \mid z; \theta) dx_{\bar{s}}. \quad (2.3)$$

Following Rubin (1976, 1978, 1980), we say that the sample design is *ignorable* if inferences based on the distribution (2.3) are equivalent to inferences based on the full distribution (2.2).

Conditions for ignorability are stronger for repeated sampling inferences based on the model than for likelihood-based inferences (Rubin 1976, Little 1980). However, if selection is achieved by probability sampling, then $f(\delta \mid z, x; \phi)$ is known and does not depend on the item values x. Under these circumstances the sample design is ignorable.

Note that for other forms of sampling the sampling mechanism may not be ignorable, and inference based on (2.3) may be subject to bias. The full model is then hard to specify unless exclusion from the sample is determined by a *known* mechanism, such as censoring with known censoring points. Thus probability sampling plays an important role in the context of superpopulation modeling even though the sampling distribution is not used as the basis of inference. This point has been noted in the literature (see, e.g., Scott 1977) but needs to be emphasized, since it provides the crucial argument against the belief that the modeling approach precludes the need to select units by probability sampling.

Example 2.1 Stratified random sampling with a completely observed variable. To illustrate the theory, suppose that z is a variable indicating J strata in the population, there are N_j units in stratum j, and x_{ij} is the item value for unit i in stratum j, $i = 1, \ldots, N_j$. Suppose that the model specifies

$$x_{ij} \sim \text{iid } G(\mu_j, \sigma_j^2), \quad i = 1, \ldots, N_j.$$

Then $\theta = \{(\mu_j, \sigma_j^2), j = 1, \ldots, J\}$ and the distribution of the population items is the product of the normal densities

$$f(x \mid z; \theta) = \prod_{j=1}^{J} \prod_{i=1}^{N_j} (2\pi\sigma_j^2)^{-.5} \times \exp(-(x_{ij} - \mu_j)^2/2\sigma_j^2).$$

Suppose that n_j units are selected from stratum j by simple random sampling without replacement. Then the distribution of sample selection is

$$f(\delta \mid z, x; \phi) = \binom{N_j}{n_j}^{-1}, \quad (2.4)$$

where $\binom{N_j}{n_j}$ is the binomial coefficient representing the number of choices of n_j items from N_j, viz $(N_j!) \{n_j! (N_j - n_j!)\}^{-1}$. Hence the likelihood (2.2) for the data is

$$\binom{N_j}{n_j}^{-1} \prod_{j=1}^{J} \prod_{i:\delta_i=1} (2\pi\sigma_j^2)^{-.5} \times \exp(-(x_{ij} - \mu_j)^2/2\sigma_j^2), \quad (2.5)$$

where the product over i is now restricted to sampled units ($\delta_i = 1$). The term (2.4) in the likelihood (2.5) does not affect likelihood inferences about θ and hence can be considered a constant, leaving the marginal distribution of sample items. If, on the other hand, the selection distribution depended on unobserved item variables, then this reduction would not be possible, and a model for the sampling mechanism would be necessary.

2.3 Superpopulation Models for Survey Data With Nonresponse

In order to model nonresponse, it is useful to divide the item variables x into two groups u and v, where u are observed for all sampled items and v are subject to nonresponse. The response pattern for v is described by the response-indicator matrix $r = (r_{ij})$, where $r_{ij} = 1$ if variable v_j is recorded for unit i and $r_{ij} = 0$ otherwise. The data are represented schematically in Figure 2. Sampled values of u, r, and v are denoted by u_s, r_s, and v_s, and nonsampled items by $u_{\bar{s}}$, $r_{\bar{s}}$, and $v_{\bar{s}}$, respectively. The sampled items of v_s are further divided into recorded values, v_{sr}, and missing values, $v_{s\bar{r}}$. Shaded areas in the figure represent the data (z, δ, u_s, r_s, v_{sr}), and values of the sample and response indicators are shown as blocks of ones and zeros separated by dashed lines. The units are arranged in rows so that the first m units are observed on all variables, the next $n - m$ units are sampled but incomplete, and the remaining $N - n$ are not sampled.

The diagram illustrates a special case of item nonresponse, where the observed values of v have a *monotone*

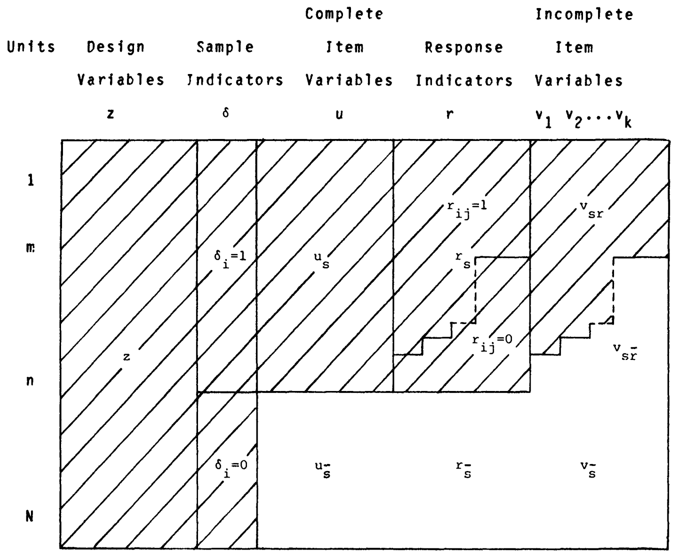

Figure 2. The Data From a Sample Survey With Monotone Nonresponse

or *nested* pattern (Anderson 1957, Hartley and Hocking 1971, Rubin 1974). That is, the set of variables v can be arranged into subsets $v_1, \dots, v_k$ such that v_j is observed for all units where v_{j+1} is observed, for $j = 1, \dots, k - 1$. In other words, v_j is *more observed* than v_{j+1} for all j. In such cases the whole data set is also monotone, with z and δ completely observed, z and δ more observed than u and r, and z, δ, u, r, more observed than v. In practice the item variables cannot always be arranged in a monotone pattern, although this may be achieved by throwing away a small amount of data (see, e.g., Marini, Olsen, and Rubin 1980). In general, it is easier to devise efficient methods for handling missing values for monotone patterns than for other patterns.

A special case of a monotone pattern is *unit nonresponse*, where $k = 1$ (that is, all partially observed variables are missing for the same units) and there are no complete item variables u. The pattern of Figure 2 then reduces to the pattern of Figure 1, with the n sampled units replaced by m responding units.

A full superpopulation model specifies a joint distribution for δ, u, r, v given z. The distribution may be specified as a product of conditional distributions of the form

$$f(\delta, u, r, v \mid z; \theta, \phi, \psi) = f(u, v \mid z; \theta) \times f(\delta \mid z, u, v; \phi) f(r \mid z, u, v, \delta; \psi). \quad (2.6)$$

The first two factors on the right side of this equation are analogous to (2.1). The last factor models the response pattern, r, through the conditional distribution of r given z, u, v, δ, indexed by a set of parameters ψ.

The distribution of the data (δ, u_s, r_s, v_{sr}) is obtained by integrating (2.6) over the missing components of u, r, and v, namely $u_{\bar s}$, $r_{\bar s}$, $v_{\bar s}$, and $v_{s\bar r}$. If we make the mild assumption that the response indicator matrices of the sampled and nonsampled units are independent given (z, u, v, δ), then this distribution takes the form

$$\begin{aligned} f(\delta, u_s, r_s, v_{sr}\, z; \theta, \phi, \psi) &= \int f(u_s, u_{\bar s}, v_{sr}, v_{s\bar r}, v_{\bar s} \mid z, \theta) \\ &\quad \times f(\delta \mid z, u_s, u_{\bar s}, v_{sr}, v_{s\bar r}, v_{\bar s}, \phi) \\ &\quad \times f(r_s \mid z, u_s, u_{\bar s}, v_{sr}, v_{s\bar r}, v_{\bar s}, \sigma; \psi) \\ &\quad \times du_{\bar s} dv_{\bar s} dv_{s\bar r}, \end{aligned} \quad (2.7)$$

where the distribution for the response indicators has been confined to the sampled units, r_s.

Most superpopulation methods for handling nonresponse are based on a model that does not include distributions for the sample and response indicators and that is restricted to the marginal distribution of the observed items u_s and v_{sr}, namely,

$$f(u_s, v_{sr} \mid z; \theta) = \int f(u_s, u_{\bar s}, v_{sr}, v_{s\bar r}, v_{\bar s} \mid z; \theta)\, du_{\bar s} dv_{\bar s} dv_{s\bar r} \quad (2.8)$$

Extending the terminology of the previous section, we say that the sample design and the response mechanism are ignorable if inferences about θ based on the distribution (2.8) are equivalent to inferences based on the full distribution (2.7). Likelihood inferences about θ based on (2.7) and (2.8) are equivalent if these expressions differ by a factor that is independent of θ. Applying the theory of Rubin (1976), sufficient conditions are as follows:

1. θ, ϕ, and ψ are distinct sets of parameters. For Bayesian inference, they are a priori independently distributed.
2. The sampling distribution $f(\delta \mid z, u, v; \phi)$ does not depend on the unobserved items $u_{\bar s}$, $v_{\bar s}$, and $v_{s\bar r}$.
3. The response distribution of sampled units $f(r_s \mid z, u, v, \delta; \psi)$ does not depend on unobserved items $u_{\bar s}$, $v_{\bar s}$, and $v_{s\bar r}$.

These conditions can be weakened for some components of θ. In particular, suppose that the joint distribution of u_s and v_s is factored as

$$f(u_s, v_s \mid z; \theta) = f(u_s \mid z; \theta_1) f(v_s \mid u_s, z; \theta_2). \quad (2.9)$$

If θ_1 and θ_2 are *distinct* parameters (or for Bayes inference, θ_1 and θ_2 are a priori independent), then the response and sampling mechanisms can be ignored for inferences about θ_1 provided condition 1 holds, together with the following two conditions (which are weaker than conditions 2 and 3):

4. The sampling distribution $f(\delta \mid z, u, v; \phi)$ does not depend on the nonsampled items $u_{\bar s}$ and $v_{\bar s}$.
5. The response distribution $f(r_s \mid z, u, v, \delta; \psi)$ does not depend on the nonsampled items $u_{\bar s}$ and $v_{\bar s}$.

In particular, response mechanisms for v_s that depend on the sampled but missing values $v_{s\bar r}$ can be ignored for inferences about θ_1, the parameters of the distribution of u_s, although they cannot be ignored for inferences about θ_2.

Conditions 2 and 4 are satisfied for probability sampling, and in most applications conditions 1 and 5 are justifiable. Hence if the parameters θ_1 and θ_2 in (2.9) are distinct, then the sample design and response mechanism can be ignored for inferences about θ_1. The crucial condition for likelihood-based inferences about θ_2 is condition 3. If this assumption is incorrect then inferences based on (2.8) are biased and it is necessary to model the response mechanism.

Example 2.2 Stratified sampling with one completely observed and one partly observed variable. Suppose that as in Example 2.1, z is a variable indicating J strata, specifically geographical regions of the population, and n_j out of N_j units are selected from stratum j by simple random sampling. There are two item variables: level of education (x_1), which is observed for all sampled units, and household income (x_2), which is missing for some sampled units. In the notation of this section, u consists of values of x_1 and v consists of values of x_2. The sampling distribution $f(\delta \mid z, u, v; \phi)$ is known and depends only on z, so conditions 2 and 4 are satisfied. For most models, conditions 1 and 5 will hold. If the response distribution of income, $f(r_s \mid z, u, v, \delta, \psi)$ (and in particular, the probability that income is observed) depends on region and education but not on the values of income itself, then condition 3 holds and the response mechanism is ignorable, although the observed income values are not necessarily a random sample of the selected units in each region. Observe that in practice nonresponse to income questions often *does* depend on the values of income, even after controlling for covariates such as region and education. Thus inferences that ignore the response mechanism are biased (Greenlees, Reece, and Zieschang 1982).

Stronger conditions are required to ignore the sampling and response mechanisms when inferences are based on repeated sampling from the superpopulation. In short, it is sufficient for inferences about θ_1 that the sampling distribution does not depend on the item values (u, v). For inferences about θ_2 we require in addition that the response distribution not depend on the nonsampled items ($u_{\bar{s}}$, $v_{\bar{s}}$) or the sampled items v_s, although it may depend on the sampled items u_s. Finally, for repeated sampling inferences involving combinations of θ_1 and θ_2 we require also that the response distribution not depend on the sampled items u_s.

Paraphrasing the conditions in this section, we can say that given probability sampling for most practical purposes the response distribution can be ignored if it does not depend on the values of items that are missing for some units. Nearly all analytic procedures for handling nonresponse effectively make this assumption. In particular, the probability of response is assumed to be constant within subclasses defined by values of design variables z or completely observed item variables, u_s.

In Section 3 we describe methods based on model distributions of the form (2.8) that assume that the response mechanism is ignorable. The methods are related to traditional sampling procedures for handling nonresponse. In Section 4 we present models for nonignorable response mechanisms that specify a distribution for the response-indicator matrix r_s given the item variables. An alternative approach is to model the conditional distribution of the observed items given the response-indicator matrix r_s, using the predictive Bayesian formulation of Rubin (1977). Although of considerable interest, this modeling approach is not discussed here.

3. LIKELIHOOD METHODS IGNORING THE RESPONSE MECHANISM

3.1 Maximum Likelihood Estimation for Incomplete Data

Before turning to specific examples, it is useful to review some general ideas concerning ML estimation for incomplete data. In principle, a likelihood analysis for incomplete data is like the analysis for complete data. A model is specified for the population values, which, if the sample design and response mechanism are ignored, induces a distribution of the form (2.8) for the data. The likelihood for the data is proportional to (2.8):

$$L(\theta; u_s, v_{sr} \mid z) \propto f(u_s, v_{sr} \mid z; \theta). \qquad (3.1)$$

Maximizing (3.1) with respect to θ yields ML estimates of the parameters. Asymptotic variances and covariances of the estimates can be derived from the inverse of the (expected) information matrix, and inferences about finite population quantities are obtained by using the model and the estimates to predict or impute the values of nonsampled or missing units.

The Bayesian analysis technically requires the additional specification of a prior distribution for the parameters. However, for many surveys the number of units is sufficiently large to minimize the influence of the prior, and hence the choice of convenient vague priors is sufficient. As noted in Section 1, the most important aspect of the formulation is the specification of a model that does not contradict the data—an area where more work is needed to satisfy practitioners (Little and Rubin 1979).

Because the data are incomplete, the likelihood tends to be complex and calculating ML estimates can be a major task. Two important computational aids are worthy of discussion—factorizations of the likelihood for monotone data patterns, discussed in the next two sections, and the expectation-maximization (EM) algorithm for arbitrary response patterns, discussed in Section 3.5.

3.2 Maximum Likelihood for Monotone Data Patterns

Factoring the likelihood was first suggested for missing data in multivariate normal problems in an article by Anderson (1957). Extensions to other distributions and more general patterns were given in Rubin (1974). Suppose first we have a single variable v subject to nonresponse. We can factor the likelihood (3.1) in the form

$$L(\theta; u_s, v_{sr} \mid z) = L(\theta_1; u_s \mid z) L(\theta_2; v_{sr} \mid u_s, z), \quad (3.2)$$

where the first component is the likelihood based on the distribution of u given z, with parameters θ_1, and the second component is the likelihood based on the distribution of v given u, z, with parameters θ_2. If the parameters θ_1 and θ_2 are *distinct*, then ML estimates $\hat{\theta}_1$ and $\hat{\theta}_2$ can be calculated independently from the factors on the right side of (3.2). By a well-known property of ML estimation, the ML estimate of any function $g(\theta_1, \theta_2)$ of θ_1 and θ_2 (for example, the mean of u given z) can be calculated by evaluating g at the ML estimates $\hat{\theta}_1$ and $\hat{\theta}_2$.

Example 3.1 Bivariate normal data. Anderson (1957) applied this idea to bivariate normal data. To relate his example to survey data, suppose z defines a single stratum, and u and v are bivariate normally distributed, with u observed for the n sampled units and v observed for $m < n$ units. The likelihood based on the joint distribution of u and v has a fairly complex form. However the factorization (3.2) simplifies the problem. The first factor corresponds to the marginal distribution of u, which is normal with mean μ_1 and variance σ_{11}. The second factor corresponds to the conditional distribution of v given u, which is normal with mean $\beta_{20} + \beta_{21} u$ variance $\sigma_{22.1}$. The first factor gives ML estimates

$$\hat{\mu}_1 = \bar{u}_n, \quad \hat{\sigma}_{11} = s_{11(n)},$$

where $\bar{u}_n$ and $s_{11(n)}$ are the mean and variance of u from the n sampled units. The second factor gives ML estimates

$$\hat{\beta}_{21} = b_{21}; \quad \hat{\beta}_{20} = \bar{v} - b_{21}\bar{u}; \quad \hat{\sigma}_{22.1} = s_{22.1},$$

where u and v are means of u and v based on the m responding units, and b_{21} and $s_{22.1}$ are the slope and residual variance from the regression of v on u, based on the responding units.

ML estimates of other parameters can be found from these estimates. For example, the mean μ_2 of v cân be written as $\mu_2 = \beta_{20} + \beta_{21}\mu_1$. Hence the ML estimate of μ_2 is the well-known regression estimate

$$\begin{aligned} \hat{\mu}_2 &= \hat{\beta}_{20} + \hat{\beta}_{21}\hat{\mu}_1 = \bar{v} - b_{21}\bar{u} + b_{21}\bar{u}_n \quad (3.3) \\ &= \bar{v} + b_{21}(\bar{u}_n - \bar{u}). \end{aligned}$$

An important point that is clear from the factored form of the likelihood is that the regression estimate is ML under more general conditions than bivariate normality (Rubin 1974). If the conditional distribution of v given u is normal with a mean that is linear in u and constant variance, then (3.3) gives the ML estimate of μ_2 provided $\bar{u}_n$ is the ML estimate of μ_1. In particular, this is the case if u is a dichotomy with a binomial distribution with mean μ_1. Generalizations of this fact to multinomial u (or z) are exploited in the examples below.

Example 3.2 Stratified estimation with unit nonresponse. Suppose that the variable u is absent, and the variable z identifies J strata in the population. Within stratum j, m_j of the n_j sampled values of v are recorded. Values of v are assumed to be iid normal random variables with mean μ_j and variance σ_j^2. The ML estimate of μ_j is simply $\bar{v}_j$, the mean of responding units in stratum j. The missing or nonsampled values of v in stratum j are estimated by $\bar{v}_j$. The resulting estimate of the population mean $\bar{V}$ of v is

$$\hat{v} = \sum_{j=1}^{J} \pi_j \bar{v}_j, \quad (3.4)$$

where π_j is the proportion of the population in stratum j, a known function of z. The value $\hat{v}$ is the ML estimate of the superpopulation mean of v. In fact, it is analogous to the regression estimate (3.3) with the interval scaled variable u replaced by the categorical design variable z. It is also the posterior mean of $\bar{V}$ given the data, under vague priors for μ_j. The posterior variance of $\bar{V}$ is

$$\operatorname{var}(\bar{V} \mid \text{data}) = \sum_{j=1}^{J} \pi_j^2 (1 - m_j/N_j) s_j^2 / m_j,$$

where s_j^2 is the sample variance of the responding units in stratum j. This expression is used to estimate the precision of the estimate of $\bar{V}$.

The estimator proposed here is analogous to the stratified estimator studied, for example, by Holt and Smith (1979). The only effect of nonresponse is to reduce the sample size in each stratum. Note that inference is conditioned on these sample sizes, and hence variation in the number of responding units is not included in the variance estimate, unlike strict randomization inferences where nonresponse is treated as another stage of probability sampling (see Sec. 3.4). The model is simplified in that it ignores heterogeneity of the v values within strata.

If the number of strata are large, or the z values indicate clusters that are not all sampled, then it may be sensible to assume that the means μ_j are themselves normally distributed with mean ν and variance τ^2. The resulting analysis is described in Scott and Smith (1969).

The previous analysis assumes that the probability of response within each stratum is not related to the value of v. The bias resulting from departures from this assumption may be reduced by further poststratifying by item variables that are thought to be related to the probability of response and to the variable of interest v. The analysis of the resulting data is the subject of the next example.

Example 3.3 Poststratification by an item variable (ctd). A categorical variable u is added to the previous example, leading to the data pattern of Example 2.2. Suppose that within stratum j, u has a multinomial distribution with

$$\Pr(u = k \mid z = j) = \pi_{k|j}.$$

Denote by (j, k) the combined stratum with $z = j$, $u = k$. Further, assume that in stratum (j, k) v is iid normal with mean μ_{jk} and variance σ_{jk}^2. The population mean V takes the form

$$\bar{V} = \sum_{j=1}^{J} \sum_{k=1}^{K} \pi_j p_{k|j} \bar{V}_{jk},$$

where $\bar{V}_{jk}$ is the population mean in stratum (j, k) and $p_{k|j}$ is the proportion of the population in stratum j. Note that $p_{k|j} \neq \pi_{k|j}$, since the latter is a superpopulation parameter in the multinomial model. The mean $\bar{V}$ is estimated by

$$\hat{v} = \sum_{j=1}^{J} \sum_{k=1}^{K} \pi_j \hat{\pi}_{k|j} \hat{\mu}_{jk}, \tag{3.5}$$

where $\hat{\pi}_{k|j}$ and $\hat{\mu}_{jk}$ are ML estimates of $\pi_{k|j}$ and μ_{jk}. Specifically,

$$\hat{\pi}_{k|j} = n_{jk}/n_{j+}, \quad \hat{\mu}_{jk} = \bar{v}_{jk},$$

where n_{j+} is the number of sampled units in stratum j, and n_{jk} and $\bar{v}_{jk}$ are the number of sampled units and the mean of v for responding units in stratum (j, k).

The estimate (3.5) corresponds to imputing for missing values of v the mean for the combined stratum, and as such relates to mean imputation, or equivalently to weighting up the values in each combined stratum by the proportion of missing values. It is well known that the naive application of standard variance formulas leads to underestimates when imputed values are present. As in the previous example, a model-based variance can be calculated that allows for the incompleteness of the data.

If the number of responding units in stratum (j, k) is small, then better estimates may be obtained by using a parsimonious model to smooth the data. In the present example, we might assume that the variables z and u are independent. Then

$$\pi_{k|j} = \pi_k \quad \text{for all } j. \tag{3.6}$$

Alternatively, if we assume that the effects of z and u on v are additive, then

$$\mu_{jk} = \mu + \alpha_j + \beta_k. \tag{3.7}$$

Furthermore, the variance σ_{jk}^2 might be assumed equal within strata or poststrata, leading to pooled estimates of variance. Models such as (3.6) and (3.7) can be fitted to the data by maximum likelihood using readily available statistical software, and resulting estimates of $\pi_{k|j}$ and μ_{jk} substituted in (3.5).

Example 3.4 Nonresponse for categorical variables. If the variable v is categorical rather than interval-scaled, then the data form a contingency table with some partially classified units, that is, units classified by z and u but not by v. The mean $\bar{V}$ is then a vector consisting of the proportion of the population in each category of v. Suppose that the values of v in stratum (j, k) have a multinomial distribution with probabilities

$$\Pr(v = l \mid z = j, u = k) = \pi_{l|jk}.$$

The marginal probability that $v = l$ is then

$$\sum_{j=1}^{J} \sum_{k=1}^{K} \pi_j \pi_{k|j} \pi_{l|jk}$$

and the estimated proportion with $v = l$ is

$$\hat{v}_l = \sum_{j=1}^{J} \sum_{k=1}^{K} \pi_j \hat{\pi}_{k|j} \hat{\pi}_{l|jk}.$$

ML estimation for models of this kind is described by Hocking and Oxspring (1974), Chen and Fienberg (1974), and Fuchs (1982). The ML estimate of $\pi_{l|jk}$ is simply the observed proportion of units in stratum (j, k) with $v = l$, and the ML estimate of $\pi_{k|j}$ is again the proportion of sampled units in stratum j with $u = k$. The resulting estimate of the vector $\bar{V}$ corresponds to allocating missing units in stratum (j, k) according to the estimated probabilities of classification $\hat{\pi}_{l|jk}$.

As in Example 3.2, the preceding procedure is only sensible with a reasonable number of responding units in each stratum (j, k). An alternative is to fit parsimonious loglinear models for the joint distribution of (z, u, v) (Goodman 1970; Bishop, Fienberg, and Holland 1975). The simplest models to fit are those that maintain the property that the parameters in each component of the factorization of the likelihood are distinct. One such model assumes that u and v are conditionally independent given z. Then

$$\pi_{l|jk} = \pi_{l|j}, \tag{3.8}$$

and is estimated by the proportion of responding units in stratum j with $v = l$. Other models, such as the one that assumes that u and z are conditionally independent given v, require more complicated fitting procedures like the EM algorithm (Fuchs 1982).

3.3 Extensions to a Set of Partially Observed Variables

Further factoring of the likelihood (3.2) is possible if there are two or more partially observed variables v that form a monotone pattern as in Figure 2. The likelihood then factors in the form

$$L(\theta_1; u_s \mid z)\, L(\theta_2; v_{sr}^{(1)} \mid z, u_s) \cdots L(\theta_{k+1}; v_{sr}^{(k)} \mid z, u_s, v_{sr}^{(1)}, \ldots, v_{sr}^{(k-1)}),$$

where $v_{sr}^{(j)}$ represents the observed data for the variables v_j. If the parameters $\theta_1, \ldots, \theta_{k+1}$ are distinct, then the analysis involves the following set of regressions:

(i) Regress u on z.
(ii) Regress v_1 on z and u
...
(k) Regress v_k on $z, u, v_1, \ldots, v_{k-1}$.

The regressions can be used to estimate population quantities by filling in missing and nonsampled values. Attention can be restricted to superpopulation parameters provided that finite population corrections are negligible.

Asymptotic estimates of precision that allow for nonresponse are relatively easy to calculate. From the Bayesian perspective, if the parameters corresponding to each regression in (i) to (k) are a priori independent, then they are a posteriori independent, and hence the posterior covariance matrix C is block diagonal. The asymptotic variance of a function g of the parameters can be obtained in the form d^TCd, where d is the vector of partial derivatives of g with respect to the parameters.

3.4 Quasi-Randomization Inferences

It is instructive to review the methods described so far from the perspective of the randomization theory. We assume probability sampling, so the sampling distribution is known and can be written

$$f(\delta \mid z, u, v; \phi) = f(\delta \mid z),$$

in the notation of Section 2. Randomization inferences from a sample of size n with complete response are based on $f(\delta \mid z)$, with item values u, v fixed. The main tasks are (a) to find a statistic T that estimates the population quantity of interest with a bias of at most order $1/n$, and (b) to find a statistic that estimates the variance of T with a bias of at most order $1/n$.

With nonresponse and $m < n$ observed item values, the natural extension is to base inferences on the joint distribution of and the response indicators r, again with u, v fixed. Scheuren (1980) coins the appropriate term *quasi randomization* to describe this mode of inference, the prefix "quasi" reflecting the fact that model assumptions are necessary for the distribution of r, since unlike the distribution of δ it is not under the control of the sampler.

The basic elements of the quasi-randomization approach are (a) a model distribution for r, which is implicit in most applications, (b) a statistic T that estimates the population quantity of interest with a bias of at most order $1/m$, and (c) a statistic that estimates the variance of T consistently.

The estimators (3.4) and (3.5) meet the criterion (b), under appropriate assumptions about the distribution of r. The former statistic is unbiased under the assumption that

$$f(r \mid z, v, \delta; \psi) = f(r \mid z, \delta; \psi).$$

That is, conditional on δ and z the response distribution (and in particular the probability of response) does not depend on values of v. The estimator (3.5) is approximately unbiased under the milder assumption that

$$f(r \mid z, u, v, \delta; \psi) = f(r \mid z, u, \delta; \psi),$$

where the response distribution is allowed to depend on the design variables z and the poststratifying variables u.

The variance of (3.4) or (3.5) under repeated sampling of r and δ can be estimated. Two issues arise here: First, the variance depends on the unknown parameters ψ, which in most cases are the probabilities of response in subclasses formed by values of z (and u); hence these parameters need to be estimated, Second, the observed proportions responding in each subclass vary in repeated sampling from the distribution of r. The question arises whether this component of variance should be included, or whether the variance of the estimates should condition on the observed number of responding units in each subclass, as in the Bayesian/likelihood approach. This question illustrates ambiguity about the appropriate degree of conditioning, which is a recurring problem in the randomization theory (Holt and Smith 1979).

The estimators (3.4) and (3.5) correspond to imputing means within subclasses of the sample. Another commonly adopted approach in surveys is to substitute observed values in the sample, as in hot-deck procedures, or external estimates, as in cold-deck methods. Such methods are not fully efficient, but can be practically convenient. For recent discussions of the hot deck, see Ford (1980) and Herzog and Rubin (1980). The latter reference extends the principle of hot deck to multiple imputations and provides a justification of the method as an approximation to Bayesian methods.

The model-based estimators (3.3), (3.6), (3.7), and (3.8) satisfy criterion (b) only under highly restrictive assumptions about the distribution of r. For example, the regression estimator (3.3) is unbiased to order $(1/m)$ in repeated sampling from δ and r only if the probability that v is recorded is the same for all values of u. The data are then analogous to data arising in double sampling (Cochran 1963, Ch. 12). If this condition is not satisified, or in other words the distribution of u is different for the respondent and nonrespondent groups, then a model is required to relate the regressions of v on u in the two groups. The latter case is important since the regression estimator serves to reduce nonresponse bias if the model is true. Finally, the estimators derived from nonignorable models in Section 4 clearly lie outside the scope of quasi-randomization inference.

The modeling approach is thus more flexible than quasi randomization for handling nonresponse for a single-item variable. This added flexibility is even more apparent when multivariate patterns of nonresponse are considered, as in the next section.

3.5 General Patterns of Nonresponse: The EM Algorithm

Convenient factorizations of the likelihood do not generally exist for data that do not have a monotone pattern. Iterative methods are then often required to calculate ML estimates. A particularly useful general method in this context is the expectation-maximization (EM) algorithm. The algorithm was first suggested in the context of particular problems (e.g., Hartley 1958). Orchard and Woodbury (1972) noted the general applicability of the method, and the ubiquity of the algorithm was amply demonstrated by Dempster, Laird, and Rubin (1977), who called it the EM algorithm since it involves an expectation step and a maximization step in each iteration.

The aim of the method is to relate estimation to the log-likelihood of the hypothetical complete data. Let $l(\theta; p, m)$ denote this log-likelihood, where θ are the parameters, p represents the present data, and m the missing data. The E step consists in integrating $l(\theta; p, m)$ over the conditional distribution of m given p, given current estimates θ_A of the parameters. The M step consists in maximizing the integrated log-likelihood with respect to θ. The resulting estimates replace the current estimates in the next E step. Under fairly general conditions the sequence of E and M steps converges to the ML estimate of θ.

Two characteristics of the algorithm are worth noting. For many important models the M step of the algorithm corresponds to a complete data calculation with missing values filled in. The E step is often a sophisticated form of imputation and as such provides a link between ML estimation and other imputation procedures for handling nonresponse.

An important model for interval scaled variables v assumes that the vector of values v_i for unit i have a multivariate normal distribution with mean

$$\beta_0 + \beta z_i + \Gamma u_i$$

and constant covariance matrix Σ. The EM algorithm for this multivariate linear regression model with arbitrary patterns of missing values is described elsewhere (Orchard and Woodbury 1972, Beale and Little 1975). An interesting aspect is that the E step of the algorithm involves imputing for a missing value v_{ij} of variable v_j for unit i the conditional mean of v_j given z_i, u_i, and the variables v_{ik} $(k \neq j)$ present for unit i. This mean is the predicted value from the regression of v_j on z, u and v variables present for unit i, calculated from current estimates of the parameters. Thus, as for monotone patterns, ML estimation involves substituting regression estimates of the missing values. The E step also involves an adjustment to the covariance matrix to allow for bias introduced by the imputation.

The multivariate normal assumption is not as restrictive as one might suppose. The important consequence of the assumption is that the regressions of the missing values are linear and additive in the observed values of v for each unit. However, nonlinear terms and interactions between the u and z variables can be included in the model without difficulty, since these variables are treated as fixed.

Another model conveniently handled by the EM algorithm occurs when the z, u, and v variables are categorical. The resulting data consist of a multiway contingency table with partially classified units. Since the z and u variables are treated as fixed, loglinear models that fix the (z, u) margins of the data are appropriate. The E step consists in allocating fractions of each partially classified unit to the cells of the table that match on the observed z, u, and v variables for that unit. The fractions are the conditional probabilities of classification given the available z,'u, and v variables and current estimates of the parameters. The M step applies a complete data ML algorithm, such as iterative proportional fitting (Darroch and Ratcliff 1972), to the data with partially classified units allocated in this way (Fuchs 1982). We illustrate the procedure with a numerical example (Little 1980).

Example 3.5 Partially classified data on two variables. For simplicity we confine attention to a single stratum (j, k) of z and u and suppose that 478 units are classified by two item variables v_1 and v_2. Of these, 300 are classified by v_1 and v_2, 90 are classified by v_1 but not v_2, and 88 are classified by v_2 but not v_1. The data are displayed in Table 1.

Figure 3 shows the steps of the EM algorithm fitting a separate parameter for the probability of falling in each cell. In the first step the cell probabilities are estimated from the completely observed units and are then used to allocate the partially classified units as indicated. For instance, the 28 partially classified units with $v_2 = 1$ have $v_1 = 1$ with probability 100/(100 + 75) and $v_1 = 2$ with probability 75/(100 + 75). Thus (28)(100)/175 = 16 are allocated to $v_1 = 1$ and 12 to $v_1 = 2$. In the next step new probabilities are found from the completed data and the procedure iterates to convergence. Final probabilities of classification are

$$p_{11} = .28, \quad p_{12} = .17, \quad p_{21} = .24, \quad p_{22} = .31 \quad (3.8)$$

4. MODELS FOR NONIGNORABLE RESPONSE PATTERNS

Models in the previous section assume that the response distribution does not depend on the values subject to nonresponse. However, in practice it is not difficult to motivate models where this is not the case. For example, if the single item variable v = Income is not recorded for some sampled individuals, then the probability of response is likely to be related to the respondent's income. The degree of dependence may be reduced if another variable such as u = Age is used as a covariate, but a residual dependence will probably exist.

Methods that introduce covariate information tend to be statistically efficient when the response mechanism is unrelated to the item variables. Much of the statistical literature of these methods is directed at demonstrating

Table 1. A 2 × 2 Table With Data Partially Classified on Both Variables

	Data Classified by							
	v_1 and v_2				v_1		v_2	
	v_2						v_2	
	1	2	Total			1	2	Total
1	100	50	150	1	30	28	60	88
v_1				v_1				
2	75	75	150	2	60			
Total	175	125	300	Total	90			

		Estimated Probabilities			Fractional Allocation of Units		
		v_2			v_2		
Step 1	v_1	1	2	v_1	1	2	
	1	100/300	50/300	1	$100+20^a+16^c$	$50+10^a+24^d$	30^a
	2	75/300	75/300	2	$75+30^b+12^c$	$75+30^b+36^d$	60^b
					28^c	60^d	
Ste		136/478	84/478		100+18.6+15.1	50+11.4+22.4	
		117/478	146/478		75+27.2+12.9	75+32.8+37.6	
Step		.28	.18		100+18.4+15.1	50+11.6+21.9	
		.24	.30		75+26.5+12.9	75+33.5+38.1	
Step		.28	.17				
		.24	.31				

Note. The superscripts in the top right hand panel indicate the allocation of the partially classified margins among the cells of the table. For example, of the 28 units with $v_2=1$ (superscript c), 16 are allocated to $v_1=1$ and 12 are allocated to $v_1=2$.

Figure 3. The EM Algorithm for Data in Table 1, Assuming Response Mechanism is Ignorable

this fact. However, the methods are even more useful when the response mechanism depends on the item variables, because of their ability to reduce nonresponse bias. The point is illustrated numerically in Example 4.2.

If the response mechanism is not ignorable, then the only way to eliminate the bias completely is to base the analysis on a model that correctly represents the response mechanism. A basic difficulty is that such models are highly sensitive to misspecification error. In some respects the problem is similar to that encountered in econometric structural equation models where parameters have to be assumed equal to zero to identify the system. If anything, the problem is more subtle for nonresponse models, as the following simple example indicates.

Example 4.1 Nonresponse for a univariate normal sample. Suppose that v is log(Income) for individual i within a given age group, and v_i is iid normal with mean μ, variance σ^2. Let $r_i = 1$ or 0 as v_i is recorded or missing, respectively, and suppose that r_i given v_i are independent Bernoulli random variables with means

$$\Pr(r_i = 1 v_i) = \Phi(\psi_0 + \psi_1 v_i), \tag{4.1}$$

where Φ is the cumulative standard normal distribution function. If $\psi_1 = 0$, the distribution of r is independent of the item values and the nonresponse mechanism is ignorable. If $\psi_1 \neq 0$, we have a form of stochastic censoring where the probability of response is a monotonic function of the item values.

Example 4.1 is a simple case of the model specified in different but equivalent terms by Heckman (1976) and Nelson (1977). Heckman introduces an unobserved variable λ that controls the response mechanism. Suppose that (v_i, λ_i) are bivariate normal with means $(\mu, 0)$, variances $(\sigma^2, 1)$, and covariance β, and that v_i is observed if and only if $\lambda_i < c$, for some unknown constant c. Then the probability of response is given by (4.1) with

$$\psi_0 = (c + \beta\mu/\sigma^2)(1 - \beta^2/\sigma^2)^{-1/2},$$

$$\psi_1 = -\beta\sigma^{-2}(1 - \beta^2/\sigma^2)^{-1/2}.$$

Note that the response mechanism is ignorable if and only if v and λ are uncorrelated (that is, $\beta = 0$).

The expected value of $\bar{v}$, the mean of the observed values of v under this model is $\mu - \beta\gamma(c)$, where $\gamma(c) = \phi(c)/\Phi(c)$ and $\phi(c)$ is the standard normal density function. Hence $\bar{v}$ is a biased estimate of μ unless $\psi_1 = \beta = 0$. The effect of censoring is illustrated in Table 2. Stem and Leaf plots of three sets of data are presented: (a) 100 values of v with $\mu = 0$, $\sigma^2 = 1$; (b) a subsample of the data in (a) with probability of inclusion given by (4.1) with $c = \beta = 0$; (c) a different subsample with c

Table 2. Stem and Leaf Plots of Distribution of Standard Normal Sample With Stochastic Censoring

	Uncensored Sample *n = 100*	*Ignorable Censoring Mechanism* *pr (y_i observed) = 0.5. m = 52*	*Nonignorable Censoring Mechanism* *pr (y_i observed) = Φ (−2.05 y_i). m = 53*
−3.5	7		7
−3			
−2.5	8		8
−2			
−1.5	57889	578	57889
−1	001111222233	1112233	001111222233
−.5	5556666778888899999	566788899999	555666778888899999
−0	0112222223344	011234	0112222234
+0	0011222222233344444	0122222234	012224
+1/2	56777778899	677789	
+1	0011113444	11144	
+1 1/2	56778	6	
+2	023	02	
+2 1/2			
+3	3		
	Sample Mean = −.03	Sample Mean = −.11	Sample Mean = −.81

$= 0$, $\beta = .9$. In both (b) and (c) the unconditional probability of response is one-half, and about half of the original sample are observed. In case (b) the response mechanism is ignorable, and the sample mean $-.11$ is a consistent estimate of the mean of the uncensored distribution, namely, zero. In case (c) the response mechanism is not ignorable. The observed distribution is skewed and the sample mean is no longer a consistent estimate of μ.

The ML estimate of μ under the model takes the form

$$\hat{\mu} = \bar{v} + \left(\frac{n - m}{m}\right) \hat{\beta}\phi(\hat{c})(1 - \Phi(\hat{c}))^{-1}, \qquad (4.2)$$

where $\hat{\beta}$ and $\hat{c}$ are ML estimates of β and c, n is the number of sampled units, and m is the number of responding units. The last term in (4.2) represents the correction for selection bias and is based entirely on the skewness in the observed values of u. Thus *the correction is totally dependent on the assumption of symmetry in the uncensored distribution of v.* If instead the original distribution were skewed and the skewness were absent in the observed data because of stochastic censoring, then the model would incorrectly infer the absence of nonresponse bias from the symmetry of the observed values. This point has been noted by Rubin (1978) and Morris (1979). The potential hazards of misspecification error suggest that a variety of nonignorable models need to be fitted as part of an assessment of sensitivity to nonresponse bias, rather than using a single model to derive unique estimates.

The introduction of a continuous latent variable λ, although not strictly necessary, is convenient for extensions to multivariate data, as in the next example.

Example 4.2 A nonignorable model for the data in Example 3.1. Suppose that (x_1, x_2, x_3) have a trivariate normal distribution with mean vector $(\mu_1, \mu_2, 0)$ and covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & 1 \end{bmatrix},$$

and a random sample $[(x_{1i}, x_{2i}, x_{3i}),\ i = 1, \ldots, n]$ is obtained from this distribution. The values of x_1 are observed for all units, the values of x_2 are observed when $x_3 < c$, for some unknown threshold c, and the values of x_3 are not observed.

To relate the model to the data of Example 3.1, restrict attention to one stratum $z = j$, and let $x_1 = u$, a completely observed item variable, and $x_2 = v$, the variable subject to nonresponse. The variable x_3 corresponds to λ in the previous example. The conditional distribution of x_3 given x_1 and x_2 is normal with mean

$$E(x_3 \mid x_1, x_2) = \beta_{31.12}(x_1 - \mu_1) + \beta_{32.12}(x_2 - \mu_2),$$

and variance $\sigma_{33.12}$, where $\beta_{31.32}$, $\beta_{32.12}$, and $\sigma_{33.12}$ are known functions of Σ. Let $r_i = 1$ if x_{2i} is observed and 0 otherwise. Then given x_{1i} and x_{2i}, r_i has a Bernoulli distribution with mean

$$\Pr(r_i = 1 \mid x_{1i}, x_{2i}) = \Pr(x_{3i} < c \mid x_{1i}, x_{2i}) = \Phi(\psi_0 + \psi_1 x_{1i} + \psi_2 x_{2i}),$$

where

$$\psi_0 = (c + \beta_{31.12}\mu_1 + \beta_{32.12}\mu_2)\sigma_{33.12}^{-1/2},$$

$$\psi_1 = -\beta_{31.12}\sigma_{33.12}^{-1/2},$$

and

$$\psi_2 = -\beta_{32.12}\sigma_{33.12}^{-1/2}.$$

We compare two estimates of μ_2: $\bar{x}_2$, the mean of $\bar{x}_2$ for the responding units, and $\hat{\mu}_2$, the regression estimate given in (3.3). Little (1980) shows that $\bar{x}_2$ has bias

$$E(\bar{x}_2 \mid \text{data}) - \mu_2 = -\beta_{23.3}\,\gamma(c) \qquad (4.3)$$

and $\hat{\mu}_2$ has approximate bias

$$E(\hat{\mu}_2 \mid \text{data}) - \mu_2 = -\beta_{23.13}\,\gamma(c), \qquad (4.4)$$

where $\gamma(c) = \phi(c)/\Phi(c)$, $\beta_{23.3}$ is the coefficient of the regression of x_2 on x_3, and $\beta_{23.13}$ is the coefficient of x_3 in the regression of x_2 on x_3 and x_1. Thus the regression estimator is effective in reducing the selection bias when $|\beta_{23.23}/\beta_{23.3}|$ is small.

A small simulation study demonstrates this property numerically. Thirty data sets of 200 units on x_1, x_2, x_3 were generated with means $\mu_1 = \mu_2 = 0$, variances $\sigma_{11} = \sigma_{22} = 1$, correlations of x_1 and x_2

$$\rho_{12} = -.9,\ -.6,\ 0,\ .6,\ .9,$$

and regression coefficients

$$(\beta_{31.12}, \beta_{32.12}) = (.0), (.6, 0), (.9, 0), (.6, .6).$$

Values of x_2 were treated as missing if $x_3 < 0$, a threshold that results in half of the values being missing on average. Table 3 tabulates $\bar{x}_2$ and $\hat{\mu}_2$ for each data set, together with the hypothetical means without nonresponse ($\bar{x}_{2n}$) and the biases predicted by (4.3) and (4.4).

For problems 1 to 6, $\beta_{32.12} = \beta_{32.12} = 0$ and the response mechanism is unrelated to x_1 or x_2. Both $\hat{\mu}_2$ and $\bar{x}_2$ are unbiased, and $\hat{\mu}_2$ has a slightly smaller variance if ρ_{12} is nonzero. The gain in precision is not apparent in the table since only one data set is generated for each problem. For problems 7 to 15, $\beta_{32.12} = 0$, and hence the conditional probability of response depends on x_1 but not on x_2. Note that $\hat{\mu}_2$ is unbiased but $\bar{x}_2$ is biased unless $\rho_{12} = 0$, with a bias that increases with $\rho_{12}{}^2$ and $\beta_{31.12}{}^2$. The observed means show the bias-correcting properties of the regression estimate in this situation.

In problems 16 to 25 the probability of response depends on x_2 but not on x_1. Here both estimators are biased, but $\hat{\mu}_2$ still reduces the bias unless $\rho_{12} = 0$. Problems 26 to 29 illustrate cases in which the partial association between x_2 and x_3 given x_1 is greater than the marginal association, and as expected the regression estimate performs worse than the sample mean. Such cases

Table 3. Estimates of the Mean of x_2 for Selected Nonresponse Mechanisms, and Their Predicted Bias

Coefficients of x_3 on x_1, x_2 ($\beta_{31 \cdot 12}$, $\beta_{32 \cdot 12}$)		Correlation of x_1, x_2 = ρ_{12}									
		−.9		−.6		.0		.6		.9	
		Estimate	Bias	Estimate	Bias	Estimate	Bias	Estimate	Bias	Estimate	Bias
(0,0)	Problem	1		2		3		4		5	
	$\bar{x}_{2n}$	.03	0	.13	0	.07	0	−.05	0	−.09	0
	$\bar{x}_2$	−.01	0	.26	0	.08	0	.06	0	−.01	0
	$\hat{\mu}_2$	.09	0	.20	0	.08	0	.03	0	−.08	0
									0		
(.6,0)	Problem	6		7		8		9		10	
	$\bar{x}_{2n}$	.02	0	−.09	0	−.05	0	.06	0	.02	0
	$\bar{x}_2$	.54	.43	.10	.29	.06	0	−.30	−.29	−.38	−.43
	$\hat{\mu}_2$	.13	0	−.26	0	.04	0	.01	0	.01	0
(.9,0)	Problem	11		12		13		14		15	
	$\bar{x}_{2n}$	0.1	0	−.09	0	−.04	0	−.05	0	−.11	0
	$\bar{x}_2$	.65	.65	.32	.43	−.02	0	−.48	−.43	−.75	−.65
	$\hat{\mu}_2$	−.04	0	−.13	0	.01	0	−.26	0	−.11	0
(0,.6)	Problem	16		17		18		19		20	
	$\bar{x}_{2n}$	−.02	0	.01	0	.05	0	−.06	0	.10	0
	$\bar{x}_2$	−.45	−.48	−.44	−.48	−.33	−.48	−.53	−.48	−.37	−.48
	$\hat{\mu}_2$	−.11	−.13	−.27	−.35	−.33	−.48	−.45	−.35	.07	−.13
(0,.9)	Problem	21		22		23		24		25	
	$\bar{x}_{2n}$	.10	0	.04	0	−.02	0	.06	0	.08	0
	$\bar{x}_2$	−.63	−.72	−.77	−.72	−.73	−.72	−.72	−.72	−.69	−.72
	$\hat{\mu}_2$	−.18	−.40	−.63	−.65	−.73	−.72	−.62	−.65	−.26	−.40
(.6,.6)	Problem	26		27		28		29		30	
	$\bar{x}_{2n}$	−.01	0	−.04	0	−.04	0	.00	0	−.01	0
	$\bar{x}_2$	−.08	−.14	−.26	−.34	−36	−.48	−.67	−.53	−.81	−.55
	$\hat{\mu}_2$	−.13	−.26	−.44	−60	−.53	−.75	−.49	−.60	−.22	−.26

are unlikely to occur in most applications, but they do indicate that the regression estimate is not automatically superior.

The ML estimate of μ_2 under the model used to generate the data takes the form

$$\hat{\mu}_2{}^* = \bar{x}_2 + \hat{\beta}_{21.1}(\bar{x}_1 - \bar{x}_{1n}) + n^{-1}\hat{\beta}_{32.12}\hat{\sigma}_{22.1} \sum_{r_i=0} \phi_i(1 - \Phi_i)^{-1},$$

where the first two terms represent the regression estimate ignoring the response mechanism and the third term is an adjustment analogous to that appearing in (4.2). The ML estimate of $\beta_{32.12}$ controls the size of the adjustment and is based on deviations of the data from the normal linear model for the regression of x_2 on x_1. The correction may reflect departures unrelated to the censoring process, and thus the final estimates are again sensitive to specification error.

The extension of this model to several covariates has been applied to data (e.g., Hausman and Spence 1977; Greenlees, Reece, and Zieschang 1982). The sensitivity of estimates to distributional assumptions may be reduced if the regressions of the response indicator and the dependent variable involve different sets of covariates; that is, in effect some coefficients in the combined set are set equal to zero. Further work is required to examine this question.

Example 4.3 Nonignorable response mechanisms for categorical data. As with continuous variables, nonignorable response models for categorical variables do not have an extensive literature. Pregibon (1977) discusses an interesting application with continuous and categorical variables.

A description of the EM algorithm for contingency tables with an arbitrary pattern of nonresponse is given in Little (1980). The model requires for each missing variable v_j the specification of the prior odds $p_j(c, d)$ that a unit belongs to category c rather than category d, for all pairs of categories (c, d). The E step of the algorithm distributes the partially classified units among cells that match on observed variables, as in the ignorable case described in Example 3.5. However in calculating the relative contributions to two cells, the ratio of the conditional probabilities for the ignorable case are multiplied by the prior odds for each pair of categories where the two cells differ. For example, suppose cell 1 has categories

$$(c_1, \ldots, c_r, c_{r+1}, \ldots, c_k)$$

of variables $v_1, \ldots, v_k$, and cell 2 has categories

$$(c_1, \ldots, c_r, c_{r+1'}, \ldots, c_{k'}).$$

Suppose we have a partially classified unit that could belong to cell 1 or cell 2. Then the posterior odds in favor

		Estimated Probabilities			Fractional Allocation of Units		
		v_2			v_2		
	v_1	1	2	v_1	1	2	
Step 1	1	100/300	50/300	1	$100+15^a+20.4^c$	$50+15^a+34.3^d$	30^a
	2	75/300	75/300	2	$75+20^b+7.6^c$	$75+40^b+25.7^d$	60^b
					28^c	60^d	
Step 2		.28	.21		100+12.2+20.3	50+17.7+35.1	
		.21	.29		75+16.0+7.7	75+44.0+24.9	
Step 3		.28	.22				
		.21	.30				

Note. The superscripts in the top right hand panel indicate the allocation of the partially classified margins among the cells of the table. For example, of the 28 units with $v_2=1$ (superscript c), 20.4 are allocated to $v_1=1$ and 7.6 are allocated to $v_1=2$.

Figure 4. The EM Algorithm for Data in Table 1, Assuming $v_1 = 1$ is Twice as Likely as $v_1 = 2$ and $v_2 = 2$ is Twice as Likely as $v_2 = 1$

of cell 1 rather than cell 2 are

$$F \times \frac{\Pr(v_1 = c_1, \ldots, v_r = c_r, \; v_{r+1} = c_{r+1}, \ldots, v_k = c_k \mid \text{data}; \theta_A)}{\Pr(v_1 = c_1, \ldots, v_r = c_r, \; v_{r+1} = c_{r+1'}, \ldots, v_k = c_{k'} \mid \text{data}; \theta_A)},$$

where $F = \prod_{j=r+1}^{k} p_j(c_j, c_{j'})$ represents an adjustment for nonresponse bias, "data" refers to observed characteristics of the unit, and θ_A denotes current estimates of the cell probabilities. The procedure is illustrated in Figure 4 for the data in Table 1.

We suppose that a priori units are twice as likely to be classified in category 1 of v_1 than category 2 of v_1, and half as likely to be classified in category 1 of v_2 than category 2 of v_2. In Example 3.5, the ratio of the conditional probabilities that $v_1 = 1$ and $v_1 = 2$, given $v_2 = 1$, at the first step were (100/300)/(75/300) = 4/3, and this determined the allocation of the 28 units unclassified by v_1 with $v_2 = 1$. In the present model this ratio is multiplied by the prior odds 2/1, giving posterior odds 8/3. Thus at the first E step 28(8/11) = 20.4 of the 28 units are allocated to $v_1 = 1$ and the rest of $v_1 = 2$. A similar procedure is adopted for the other cells. The final estimates of the cell probabilities are $p_{11} = .28$, $p_{12} = .22$, $p_{21} = .21$, $p_{22} = .30$, which can be compared with the results for the ignorable case in (3.8).

5. CONCLUSION

This article relates the extensive literature on incomplete data to the problem of nonresponse in sample surveys. The literature discussed here is rarely applied in the survey context, but I believe it has profound implications in the design and analysis of surveys where nonresponse is a problem. The theoretical concept of ignorability provides insight into the role of probability sampling in surveys and into the difficulties of modeling nonresponse. The theory of maximum likelihood provides a large class of efficient estimators that can be used to eliminate or reduce nonresponse bias. For special data patterns like those discussed in Section 3.2, factorizations of the likelihood often yield computationally straightforward estimates of population quantities in the presence of nonresponse, together with model-based standard errors that allow for the incompleteness of the data. For general data patterns the EM algorithm discussed in Section 4 is a valuable computational tool. Finally, models like those outlined in Section 4 can form the basis for sensitivity analyses to assess the effect of nonignorable nonresponse on survey estimates.

Much work remains to be done, both in developing useful models for survey nonresponse, and in making model-based technology available to the survey practitioner. The (in my view) healthy skepticism with which many practitioners regard model-based methods needs to be tempered by successful modeling exercises in the context of real surveys. On the other hand, more work is needed to relate model-based methods to methods currently used to handle nonresponse in surveys, such as the hot deck. I believe that the National Academy of the Sciences Panel on Incomplete Data has laid the groundwork for this activity. Finally, computational algorithms for handling nonresponse need to be made more generally available.

[*Received November 1980. Revised November 1981.*]

REFERENCES

ANDERSON, T.W. (1957), "Maximum Likelihood Estimates for a Multivariate Normal Distribution When Some Observations are Missing," *Journal of American Statistical Association*, 52, 200–203.

BEALE, E.M.L., and LITTLE, R.J.A. (1975), "Missing Values in Multivariate Analysis," *Journal of the Royal Statistical Society*. Ser. B, 37, 129–146.

BISHOP, Y.M.M., FIENBERG, S.E., and HOLLAND, P.W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, Mass.: MIT Press.

BREWER, K.R.W., and SÄRNDAL, C.E. (1979), "Six Approaches to Enumerate Survey Sampling," in *Proceedings of the Symposium on Incomplete Data*, Washington, D.C.: National Academy of Sciences.

CHEN, T., and FIENBERG, S.E. (1974), "Two-Dimensional Contingency Tables With Both Completely and Partially Classified Data," *Biometrics*, 30, 629–642.

COCHRAN, W.G. (1963), *Sampling Techniques*, 3rd ed., New York: John Wiley.

DARROCH, J.N., and RATCLIFF, D. (1972), "Generalized Iterative Scaling for Log-Linear Models," *Annals of Mathematical Statistics*, 43, 1470–1480.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm" (with Discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1, 1–38.

ERICSON, W.A. (1969), "Subjective Bayesian Models in Sampling

Finite Populations, I," *Journal of the Royal Statistical Society*, Ser. B, 31, 195–234.

FAY, R.E., and HERRIOTT, R.A. (1979), "Estimates of Income for Small Places: An Application of James Stein Procedures to Census Data," *Journal of the American Statistical Association*, 74, 269–278.

FORD, B.N. (1980), "An Overview of Hot Deck Procedures," in *Non-Response in Sample Surveys: The Theory of Current Practice*, Part III, Panel on Incomplete Data, National Academy of Sciences, Washington, D.C.

FUCHS, C. (1982), "Maximum Likelihood Estimation and Model Selection in Contingency Tables With Missing Data," *Journal of the American Statistical Association*, 77, 270–278.

GOODMAN, L.A. (1970), "The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications," *Journal of the American Statistical Association*, 65, 225–256.

GREENLEES, W.S., REECE, J. S. and ZIESCHANG, K.D. (1982), "Imputation of Missing Values When the Probability of Response Depends Upon the Variable Being Imputed," *Journal of the American Statistical Association*, 77, 251–261.

HARTLEY, H.O. (1958), "Maximum Likelihood Estimation From Incomplete Data," *Biometrics*, 14, 174–194.

HARTLEY, H.O., and HOCKING, R.R. (1971), "The Analysis of Incomplete Data," *Biometrics*, 14, 174–194.

HAUSMAN, J.A., and SPENCE, A.M. (1977), "Non-Random Missing Data," Working Paper, Massachusetts Institute of Technology, Dept. of Economics.

HECKMAN, J.D. (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.

HERZOG, T.N., and RUBIN, D.B. (1980), "Using Multiple Imputations to Handle Non-Response in Sample Surveys," in *Non-Response in Sample Surveys: The Theory of Current Practice, Part III*, Panel on Incomplete Data, National Academy of Sciences, Washington, D.C.

HOCKING, R.R., and OXSPRING, H.H. (1974), "The Analysis of Partially Categorized Contingency Data," *Biometrics*, 30, 469–483.

HOLT, D., and SMITH, T.M.F. (1979), "Poststratification," *Journal of the Royal Statistical Society*, Ser. A, 142, 33–66.

LITTLE, R.J.A. (1980), "Superpopulation Models for Non-Response. I: The Ignorable Case," and "II: The Non-Ignorable Case" in *Non-Response in Sample Surveys: The Theory of Current Practice, Part V*, Panel on Incomplete Data, National Academy of Sciences, Washington, D.C.

LITTLE, R.J.A., and RUBIN, D.B. (1979), Discussion of "Six Approaches to Enumerate Survey Sampling" by K.R.W. Brewer and C.E. Särndal, in *Proceedings of the Symposium on Incomplete Data*, Washington, D.C.: National Academy of Sciences.

MARINI, M.M., OLSEN, A.R., and RUBIN, D.B. (1980), "Maximum Likelihood Estimation in Panel Studies With Missing Data," in *Sociological Methodology 1980*, San Francisco: Jossey-Bass.

MORRIS, C.N. (1979), "Non-Response Issues in Public Policy Experiments, With Emphasis on the Health Insurance Study," in *Proceedings of the Symposium on Incomplete Data*, Washington, D.C.: National Academy of Sciences.

NELSON, F.D. (1977), "Censored Regression Models With Unobserved, Stochastic Censoring Thresholds," *Journal of Econometrics*, 6, 581–92.

ORCHARD, T., and WOODBURY, M.A. (1972), "A Missing Information Principle: Theory and Applications," in *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 697–715.

PREGIBON, D. (1977), "Typical Survey Data: Estimation and Imputation," *Survey Methodology*, 2, 70–102.

ROYALL, R.M. (1970), "On Finite Population Sampling Theory Under Certain Linear Regression Models," *Biometrika*, 57, 377–387.

ROYALL, R.M., and HERSON, J. (1973), "Robust Estimation From Finite Populations," *Journal of the American Statistical Association*, 68, 880–889.

RUBIN, D.B. (1974), "Characterizing the Estimation of Parameters in Incomplete Data Problems," *Journal of the American Statistical Association*, 69, 467–474.

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

——— (1977), "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–543.

——— (1978), "Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse" (with discussion and reply), in *Imputation and Editing of Faulty or Missing Survey Data*, 1–9 U.S. Social Security Administration and Bureau of the Census.

——— (1980), "Conceptual Issues in the Presence of Nonresponse," *Nonresponse in Sample Surveys: The Theory of Current Practice*, Part III, Panel on Incomplete Data, National Academy of Sciences, Washington, D.C.

SCOTT, A.J. (1977), "On the Problem of Randomization in Survey Sampling," *Sankhya*, Ser. C, 39, 1–9.

SCOTT, A.J., and SMITH, T.M.F. (1969), "Estimation in Multistage Surveys," *Journal of the American Statistical Association*, 64, 830–840.

SCHEUREN, F. (1980), "Weighting Adjustments for Unit Non-Response," in *Non-Response in Sample Surveys: Theory of Current Practice*, Part III, Panel on Incomplete Data, National Academy of Sciences, Washington, D.C.