Journal of Official Statistics
Vol. 7. No. 2, 1991, pp. 167–179

Nonresponse Imputation with Multiple Sources of Nonresponse

R. L. Hinde[1] and R. L. Chambers[1]

Abstract: This paper extends recent developments in imputation techniques which allow for sample design effects as well as non-ignorable nonresponse. A model which incorporates multiple sources of nonresponse is developed and a case study on data from a survey of farms is presented.

Key words: Design adjustment; poststratification; probit analysis; EM algorithm; selection models; sample survey; regression imputation.

1. Introduction

This paper further develops the regression based approach to imputation for non-ignorable item nonresponse in finite population samples that was described in Chambers (1988). In particular, improvements and refinements to the algorithm in that reference are presented, as well as an extension to allow for multiple sources of nonresponse. Throughout, our emphasis is on imputation in order to improve point estimation of population quantities. The related issue of imputation as a method of recovering population variability, and hence improved coverage properties for interval estimates, is left for another time.

In the estimation of population quantities it is not always necessary to impute for individual nonrespondents. A class of estimation schemes which adjust for nonresponse without imputation are those which modify the sample weights (Oh and Scheuren 1983). A special case of this approach is to simply ignore the nonrespondents and treat the respondents as the full sample. Such approaches tend to be arithmetically equivalent to imputation schemes, so that they can be considered to impute implicitly.

Several techniques for item imputation exist. One can substitute external or previous information (Platek and Gray 1983). Bayesian methods have been developed (e.g., Chiu and Sedransk (1986) for small sample surveys) but these have not had a great effect in this field. Hot-decking (Chapman 1976; Ford 1983) involves judicious selection of a respondent's value for the imputed value. While this technique is generally satisfactory and quite widely used, it is difficult to assess analytically. In hot-decking it is presumed that the variable of interest is related to other variables known for each unit, for example, stratum or size variables. Imputation is achieved by choosing an imputed value from among the values of the respondents whose predictor

[1] Australian Bureau of Agricultural and Resource Economics, GPO Box 1563, Canberra, ACT 2601, Australia.

variables are in some sense similar to those of the nonrespondent. Regression imputation (Little 1983) works on the same presumption. A regression model is fitted to the respondents' data and imputed values for the nonrespondents are obtained by using their known predictor variables in the fitted model.

In applying regression imputation to sample nonresponse, two problems need to be addressed. Firstly, one has to decide whether the regression model used in imputation should reflect sample characteristics or whether it should reflect population characteristics. In the latter case, sample design effects need to be allowed for by using design adjusted methods to fit this model (Chambers 1986). The second problem occurs if the nonresponse mechanism is not ignorable (Little 1983; Little and Rubin 1987; Rubin 1987), that is, where the probability of nonresponse is related to the value of the variable of interest. The use of non-ignorable nonresponse models as well as design adjustment for regression imputation were considered in Chambers (1988) and this paper continues with this approach.

Throughout this paper we assume that nonresponse can be characterised by the operation of a selection mechanism (Heckman 1979). That is, the probability of a nonresponse depends on the underlying data value that we are attempting to measure. Selection mechanisms are particularly suited to the economic variables with which we are mainly concerned. In particular, those variables for which nonresponse is a problem are continuous and become increasingly "sensitive" to public disclosure as they either increase or decrease. With the additional assumption of underlying normality for the regression residuals, such models have the further advantage of being estimable from the respondent data.

Balanced against this, however, is the fact that it is effectively impossible to test how well the assumption of a normal selection mechanism fits the observed nonresponse, since the data required to verify such an assumption (the nonrespondents' data) are, by definition, unavailable. If the nonresponse is small, the distribution of nonrespondents may not differ markedly from that of all cases, and correct identification of the nonresponse mechanism is not an important issue. However, if the nonresponse is extensive there is the possibility that the nonnormality induced by a selection mechanism can become confounded with misspecification of the regression model itself (Little and Rubin 1987). In such cases, the analyst should take some care when specifying a model for the nonresponse process, and other mechanisms for nonresponse should also be considered, for example, a mixture mechanism where each individual is first classified as a respondent or nonrespondent, and data values are then generated conditional on this status. Models which incorporate such a mixture mechanism are easy to work with. Unfortunately, they suffer from the drawback that they are non-estimable (without access to the nonrespondents' data) unless the nonresponse is assumed to be ignorable or a Bayesian approach is taken with appropriately specified priors for the parameters underlying the distribution of the nonrespondent data (Rubin 1987).

In the situation considered in this paper, a follow-up survey of nonrespondents is not possible. Given the essential non-identifiability of the underlying nonresponse mechanism from the respondents' data, our approach therefore is to focus on a particular class of models for the nonresponse (the normal selection models) which seem subjectively suited to the variables of concern to us and to empirically evaluate their per-

formance. Alternative nonresponse models are therefore not considered any further in this paper.

In Section 2.1 we introduce the basic model for the item to be imputed. This consists of a standard linear regression linked to a selection mechanism that allows for non-ignorable nonresponse. In Sections 2.2 and 2.3 respectively, we develop within stratum and across stratum applications of this model. In both, the selection component of the model is fitted by a probit analysis while the regression component equations are solved via the EM algorithm. Section 2.4 compares these approaches to more standard regression based imputation methods which treat the nonresponse as ignorable. Section 2.5 extends the methods developed in Sections 2.2 and 2.3 to the case of multiple sources of nonresponse. Section 3 provides some empirical evidence for the performance of these methods by applying them to survey data containing multiple sources of nonresponse. Finally, Section 4 provides some discussion on issues in nonresponse imputation.

2. Derivation of Imputation Strategies

2.1. *Preliminaries*

Assume a finite population made up of $i = 1, \ldots, N$ units which can be split into $h = 1, \ldots, H$ relatively homogeneous subpopulations or strata, each of size N_h. Let $i \in h$ denote a unit in subpopulation h. For each i define a dependent variable Y_i and a column vector X_i of p associated explanatory variables. Put $Z_i = (1, X_i')'$, where $'$ denotes transpose, and let Y denote the population column vector $(Y_1, \ldots, Y_N)'$ and Z the corresponding $N \times (p + 1)$ matrix $(Z_1, \ldots, Z_N)'$. Nonresponse is assumed to be a problem only with the variable Y_i. Throughout this paper the subscript 0 will denote restriction to sample respondents, 1 to sample nonrespondents and r to non-sample units, while s will denote the complete set of n sampled units. Expressions such as Y_{h0}, Z_{hs}, and n_{h0} then follow.

Given the assumption of within stratum homogeneity, the vectors $(Y_i, X_i')'$, $i \in h$ will be assumed to be independently and identically distributed, with

$$E\begin{bmatrix} Y_i \\ X_i \end{bmatrix} = \begin{bmatrix} \mu_{hY} \\ \mu_{hX} \end{bmatrix} \text{ and}$$

$$\operatorname{cov}\begin{bmatrix} Y_i \\ X_i \end{bmatrix} = \begin{bmatrix} \sigma_{hYY} & \sigma_{hXY}' \\ \sigma_{hXY} & \sigma_{hXX} \end{bmatrix}. \tag{1}$$

We further assume that, for each unit $i \in h$, conditioning on X_i leads to

$$Y_i | X_i = Z_i'\beta_h + U_i \tag{2}$$

where the U_i are independent $\mathfrak{N}(0, \sigma_h^2)$ random variables. Note that if (1) is extended to assume multivariate normality, (2) then follows, with β_h and σ_h^2 functions of the parameters in (1).

The nonresponse component of the model assumes that there exists, for every unit, a fixed but unknown threshold value C_i and an unobservable random variable V_i, such that unit i responds for Y_i whenever $V_i \leqslant C_i$. These V_i are assumed to be independent $\mathfrak{N}(0, 1)$ random variables. In general, V_i and Y_i may be related, so the nonresponse is potentially non-ignorable. We model this link by assuming that, for $i \in h$

$$\operatorname{cov}(Y_i, V_i | X_i) = \omega_h \geqslant 0$$

and

$$C_i = W_i'\lambda$$

where W_i is a vector of q known predictor variables, and λ is an unknown parameter vector. Because V_i is normally distributed, it follows that the probability that $V_i \leqslant C_i$ is $\Phi(W_i'\lambda)$, where Φ denotes the standardized

normal distribution function (and ϕ will later be used to denote the corresponding standardized normal density function). The vector W_i may or may not contain some of the variables that occur in X_i.

2.2. *Within stratum imputation*

In this section we develop an imputation strategy for the missing data vector Y_{h1} that depends only on information from the sample in stratum h. The underlying regression model used in imputation therefore reflects only the relationship between Y_i and X_i for the sample units in this stratum. Note that this information consists of:

i. Y_{h0}

ii. Z_{hs}

iii. That $V_{h0} \leqslant C_{h0}$ holds

iv. That $V_{h1} > C_{h1}$ holds

where $\leqslant$ and $>$ denote componentwise inequality. The imputation methods described in this paper rely heavily on taking expectations conditional on this information. Estimates of these conditional expectations will be denoted $\hat{E}_{hs}(.)$ in what follows.

Item (iii), combined with the fact that cov $(Y_i, V_i|X_i) = \omega_h$ for all $i \in h$, means that the ordinary least squares estimator for β_h in (2), based only on the respondents' data, is biased when $\omega_h \neq 0$. The information (iii) and (iv) can be incorporated into the model (2) by assuming joint normality of U_i and V_i. Given X_i, Y_i and V_i are then jointly normal

$$\begin{pmatrix} Y_i \\ V_i \end{pmatrix} \Big| X_i \sim \mathfrak{N}\left(\begin{bmatrix} Z_i'\beta_h \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_h^2 & \omega_h \\ \omega_h & 1 \end{bmatrix} \right) \quad (3)$$

with independence between different $(Y_i, V_i)'$ vectors.

From this form of the model, estimates for the unknown parameters β_h, σ_h^2, and ω_h in (3) can be calculated via the EM algorithm (Dempster, Laird, and Rubin 1977; Little and Rubin 1987). In the M-step of this algorithm, maximum likelihood estimating equations for these parameters are obtained assuming Y_h, Z_h and V_h are all observed. The log-likelihood function is then

$$\begin{aligned} \log L(\beta_h, \sigma_h^2, \omega_h; Y_h, Z_h, V_h) &= -\tfrac{1}{2}N_h \log(2\pi) - \tfrac{1}{2}N_h \log(\sigma_h^2 - \omega_h^2) \\ &\quad - \tfrac{1}{2}V_h'V_h - (Y_h - Z_h\beta_h - \omega_h V_h)' \\ &\quad \times (Y_h - Z_h\beta_h - \omega_h V_h) \end{aligned}$$

and differentiation with respect to β_h yields the full information maximum likelihood estimating equation

$$Z_h'(Y_h - Z_h\beta_h - \omega_h V_h) = 0$$

with similar equations for σ_h^2 and ω_h. In the E-step the various functions of Y_h, Z_h and V_h contained in these full information estimating equations are replaced by their conditional expectations given the actual sample information (i)–(iv), and the equations solved. Since these conditional expectations will typically depend on the unknown parameters, they are calculated on the basis of initial estimates of these parameters. These two steps are then iterated until the solutions to the likelihood estimating equations above converge. After some algebra we can show that this process of iterative substitution leads to maximum likelihood estimating equations with solutions:

$$\hat{\beta}_h = (Z_{h0}'A_{h0}Z_{h0})^{-1}Z_{h0}'A_{h0}Y_{h0} \quad (4)$$

$$\hat{\omega}_h = \frac{\hat{E}_{hs}(V_{h0}')(Y_{h0} - Z_{h0}\hat{\beta}_h)}{\hat{E}_{hs}(V_{h0}'V_{h0})} \quad (5)$$

$$\hat{\sigma}_h^2 = \hat{\omega}_h^2 + Y_{h0}'A_{h0}(Y_{h0} - Z_{h0}\hat{\beta}_h)/n_{h0} \quad (6)$$

where

$$A_{h0} = I_{h0} - \frac{\hat{E}_{hs}(V_{h0})\hat{E}_{hs}(V_{h0}')}{\hat{E}_{hs}(V_{h0}'V_{h0})}$$

I_{h0} is the $n_{h0} \times n_{h0}$ identity matrix and, for $i \in s_{h0}$

$$\hat{E}_{hs}(V_i) = \hat{\omega}_h[Y_i - Z_i'\hat{\beta}_h]/\hat{\sigma}_h^2 - \hat{\varphi}_h\phi(\hat{\zeta}_i)/\Phi(\hat{\zeta}_i) \quad (7)$$

and

$$\hat{E}_{hs}(V_i^2) = [\hat{E}_{hs}(V_i)]^2 + \hat{\varphi}_h^2[1 - \hat{\zeta}_i\phi(\hat{\zeta}_i)/\Phi(\hat{\zeta}_i) - \phi(\hat{\zeta}_i)^2/\Phi(\hat{\zeta}_i)^2] \quad (8)$$

with

$$\hat{\varphi}_h = (1 - \hat{\omega}_h^2/\hat{\sigma}_h^2)^{1/2}$$

and

$$\hat{\zeta}_i = (C_i - \hat{\omega}_h[Y_i - Z_i'\hat{\beta}_h]/\hat{\sigma}_h^2)/\hat{\varphi}_h.$$

Equations (7) and (8) are derived by noting that if $i \in s_{h0}$, V_i conditioned on the stratum h sample data has a $\mathfrak{N}(\omega_h[Y_i - Z_i'\beta_h]/\sigma_h^2, 1 - \omega_h^2/\sigma_h^2)$ distribution truncated to values below C_i.

Equations (4)–(8) depend on the unknown quantities C_i of the nonresponse component of the model given in Section 2.1. These can be replaced by $\hat{C}_i = W_i'\hat{\lambda}$, where $\hat{\lambda}$ is the estimate of λ obtained after fitting a standard probit model (see Kotz and Johnson 1986), with covariates W_i, to the indicator vector for sample response and nonresponse. Thus the estimation of β_h, σ_h^2 and ω_h is done in two stages: a probit stage to obtain the $\hat{C}_i$, followed by a regression stage of iterating (4)–(8), with $\hat{C}_i$ used for C_i. Imputed values for the missing data vector Y_{h1} then follow from

$$\hat{Y}_{h1} = \hat{E}_{hs}(Y_{h1}) = Z_{h1}\hat{\beta}_h + \hat{\omega}_h\hat{E}_{hs}(V_{h1}) \quad (9)$$

where, for $i \in s_{h1}$

$$\hat{E}_{hs}(V_i) = \phi(\hat{C}_i)/[1 - \Phi(\hat{C}_i)]. \quad (10)$$

2.3. *Across stratum imputation*

The within stratum imputation scheme derived in the previous section may be unstable, especially if stratum sample sizes are small. A lower mean square error may be obtained by replacing $\hat{\beta}_h$ in (9) by a common $\hat{\beta}$ estimated across all the strata. While this would produce a higher model bias, this may be outweighed by its lower sampling variability. One approach to constructing such a $\hat{\beta}$ is via an "average" population level regression model of the form

$$E[Y_i|X_i] = Z_i'\beta.$$

That is, a model for the conditional expectation of Y_i given X_i where i is some arbitrarily chosen population unit. It can be shown (Chambers 1986; Chambers 1988), that, given (2), estimation of β in such an "average" model requires one to simultaneously adjust for differential stratum sampling fractions and the non-ignorable nonresponse mechanism operating within each of the strata. This leads to an estimator $\hat{\beta}$ of β of the form

$$\hat{\beta} = (\widehat{Z'Z})^{-1}(\widehat{Z'Y}) \quad (11)$$

where

$$\widehat{Z'Z} = \sum_h \frac{N_h}{n_h} Z_{hs}'Z_{hs} \quad (12)$$

and

$$\widehat{Z'Y} = \sum_h \frac{N_h}{n_h} (Z_{h0}'Y_{h0} + Z_{h1}'\hat{E}_{hs}(Y_{h1})). \quad (13)$$

Here $\hat{E}_{hs}(Y_{h1})$ is the vector of within stratum imputations (9). Final across stratum imputations for the vector Y_{h1} are then given by

$$\hat{Y}_{h1} = Z_{h1}\hat{\beta} + \hat{\omega}_h\hat{E}_{hs}(V_{h1}) \quad (14)$$

where $\hat{E}_{hs}(V_{h1})$ is derived from (10).

It has been pointed out by a referee that the definition of an "average" $\hat{\beta}$ via (11)

above does not represent the only way one can stabilize the imputation process. Other possibilities include fitting models of the form

$$E[Y_i|X_i;\ i \in h] = \alpha_h + X_i'\beta$$

corresponding to the assumption of common slope parameters across strata. Although we do not consider this model further here, it does represent an alternative to the within stratum model considered in Section 2.2, and would be well worth considering if the estimated slope coefficients from (4) were all "close" to one another.

2.4. *Regression imputation techniques that assume ignorable nonresponse*

If one is prepared to assume nonresponse is ignorable then the simplest imputation method is to fit (2) via ordinary least squares to the respondent data within each of the poststrata and then impute via $\hat{Y}_{h1} = Z_{h1}\hat{\beta}_{h,\mathrm{OLS}}$ where

$$\hat{\beta}_{h,\mathrm{OLS}} = (Z_{h0}'Z_{h0})^{-1}Z_{h0}'Y_{h0}. \tag{15}$$

On the other hand, if one assumes (2) holds at population rather than stratum level then standard design adjustment ideas (Chambers 1986) lead to weighted least squares (WLS) imputations defined by $\hat{Y}_{h1} = Z_{h1}\hat{\beta}_{\mathrm{WLS}}$ where

$$\hat{\beta}_{\mathrm{WLS}} = \left(\sum_h \frac{N_h}{n_{h0}} Z_{h0}'Z_{h0}\right)^{-1} \times \sum_h \frac{N_h}{n_{h0}} Z_{h0}'Z_{h0}\hat{\beta}_{h,\mathrm{OLS}}. \tag{16}$$

The imputation methods developed in Sections 2.2 and 2.3 are extensions of these ignorable nonresponse imputation methods. To see this, consider the situation where the nonresponse model developed in Section 2.1 has constant C_i values, so the nonresponse is ignorable. One would then expect the $\hat{C}_i$ also to be constant apart from random variation.

Table 1. Relationships between regression imputation procedures

Regression model	Type of nonresponse	
	Missing at random	Selection mechanism
Varies between poststrata	OLS Procedure	Within Stratum Procedure
"Average" population level model	WLS Procedure	Across Stratum Procedure

It can be shown that if the $\hat{C}_i$ are constant then the within stratum imputations in equation (9) reduce to the OLS imputations defined by (15). Thus the OLS imputations provide a convenient benchmark which can be used to assess whether the probit stage of the within stratum procedure is effective.

It can also be shown that if the $\hat{C}_i$ are constant then the across stratum imputations in (14) reduce to $\hat{Y}_{h1} = Z_{h1}\hat{\beta}$ with

$$\hat{\beta} = \left(\sum_h \frac{N_h}{n_h} Z_{hs}'Z_{hs}\right)^{-1} \sum_h \frac{N_h}{n_h} Z_{hs}'Z_{hs}\hat{\beta}_{h,\mathrm{OLS}}. \tag{17}$$

This is similar to the WLS formula (16), the difference being that (17) contains a design adjustment that allows for the known non-respondents' predictor variables Z_{h1}. Once again, the WLS procedure, given by (16), provides a benchmark which can be used to assess the effectiveness of this extra level of design adjustment and the probit stage.

The relationships between these different imputation schemes are shown in Table 1.

2.5. *Multiple sources of nonresponse*

Consider the situation where nonresponse can be categorised as due to any of K possible causes, and is non-ignorable. This can be modelled by the movement of K

independent censoring variables (V_{ik}: $k = 1, \ldots, K$) across K thresholds (C_{ik}: $k = 1, \ldots, K$). A response corresponds to all K censoring variables staying below their thresholds simultaneously while for a nonresponse to occur, at least one of the V_{ik} must exceed the corresponding C_{ik}. Keeping the same W_i to model each of the K categories of nonresponse, the model (3) now becomes, for all $i \in h$ and $k = 1, \ldots, K$

$$\begin{pmatrix} Y_i \\ V_i \end{pmatrix} \Bigg| X_i \sim \mathfrak{N}\left(\begin{bmatrix} Z_i'\beta_h \\ 0_K \end{bmatrix}, \begin{bmatrix} \sigma_h^2 & \omega_h' \\ \omega_h & I_K \end{bmatrix} \right) \quad (18)$$

and

$$C_{ik} = W_i'\lambda_k \quad (19)$$

where now $V_i = (V_{i1}, \ldots, V_{iK})'$, $\omega_h = (\omega_{h1}, \ldots, \omega_{hK})'$, 0_K is the zero vector of order K and I_K is the identity matrix of order K.

For practical reasons, we restrict ourselves to the situation where each nonrespondent is uniquely allocated to a single known nonresponse category. The information from the sample in stratum h will then include (i)–(iii) of Section 2.2, where V_{h0} is now the $n_{h0} \times K$ matrix whose rows consist of the V_i' for the respondents in stratum h and the corresponding rows of C_{h0} consist of the vectors $C_i' = (C_{i1}, \ldots, C_{iK})'$ for the same i. Let V_{hk} denote the matrix whose rows consist of the V_i' for the category k nonrespondents in stratum h, with C_{hk} defined similarly. The sample information about the V_{hk} depends upon the process determining this allocation. For example, suppose the censoring is carried out in a sequential manner, so that V_{i1} is determined prior to V_{i2}, and so on, until either a censoring event takes place, or all $V_{ik} \leqslant C_{ik}$. The sample information about the values in V_{hk} is then:

(iv) That $V_{ik} > C_{ik}$ and $V_{ij} \leqslant C_{ij}$, $j = 1, \ldots, k-1$, hold.

Notice there is no sample information about the V_{ij}, $j > k$, since it is assumed that these values have no effect on a category k nonrespondent.

The application of the EM-algorithm for K categories is a straightforward extension of the single category case. It turns out that the probability process by which nonrespondents are allocated to a unique nonresponse category has no effect on the algorithm's solution, which is

$$\hat{\beta}_h = (Z_{h0}'A_{h0}Z_{h0})^{-1}Z_{h0}'A_{h0}Y_{h0} \quad (20)$$

$$\hat{\omega}_h = (\hat{E}_{hs}V_{h0}'V_{h0})^{-1}\hat{E}_{hs}(V_{h0}')(Y_{h0} - Z_{h0}\hat{\beta}_h) \quad (21)$$

$$\hat{\sigma}_h^2 = \hat{\omega}_h'\hat{\omega}_h + Y_{h0}'A_{h0}(Y_{h0} - Z_{h0}\hat{\beta}_h)/n_{h0} \quad (22)$$

where

$$A_{h0} = I_{h0} - \hat{E}_{hs}(V_{h0})(\hat{E}_{hs}(V_{h0}'V_{h0}))^{-1} \times \hat{E}_{hs}(V_{h0}').$$

From (18) we see that

$$V_i | Y_i, X_i \sim \mathfrak{N}(\omega_h[Y_i - Z_i'\beta_h]/\sigma_h^2, I_K - \omega_h\omega_h'/\sigma_h^2).$$

Ideally, the conditional expectations of V_{h0} and $V_{h0}'V_{h0}$ in A_{h0} above are then obtained by appropriate integration of this K-dimensional distribution over that region of $\mathscr{R}^K$ defined by the restrictions $V_{i(k)} \leqslant C_{i(k)}$. In general, this will require numerical integration. First order approximations to these conditional expectations can be obtained by setting the off-diagonal entries in the conditional variance-covariance matrix of V_i above to zero. Using this approach gives (7) and (8) where V_i, $\hat{\omega}_h$, $\hat{\varphi}_h$ and $\hat{\zeta}_i$ are now $K \times 1$ vectors, and all operations on vectors are taken as being elementwise. Imputed values of Y_i for category k nonrespondents in

stratum h are then obtained from

$$\hat{Y}_{hk} = \hat{E}_{hs}(Y_{hk}) = Z_{hk}\hat{\beta}_h + \hat{E}_{hs}(V_{hk})\hat{\omega}_h. \quad (23)$$

Unfortunately, evaluation of $\hat{E}_{hs}(V_i)$ for $i \in s_{hk}$ depends on the method of allocation to the nonresponse categories. Under (iv) above, this will be, for a category k nonrespondent

$$\hat{E}_{hs}V_{ij} = \begin{cases} -\phi(\hat{C}_{ij})/\Phi(\hat{C}_{ij}) & j < k \\ \hat{\phi}(C_{ij})/[1 - \Phi(\hat{C}_{ij})] & j = k \\ 0 & j > k \end{cases} \quad (24)$$

Similarly, the across stratum imputation method of Section 2.3 can be extended to allow for multiple nonresponse categories. In particular, (11) still holds, but with

$$\widehat{Z'Y} = \sum_h \frac{N_h}{n_h}\left(Z'_{h0}Y_{h0} + \sum_k Z'_{hk}\hat{E}_{hs}(Y_{hk})\right) \quad (25)$$

where $\hat{E}_{hs}(Y_{hk})$ is the within stratum imputation (23). It follows that the across stratum imputation for category k nonresponse in stratum h is

$$\hat{Y}_{hk} = Z_{hk}\hat{\beta} + \hat{E}_{hs}(V_{hk})\hat{\omega}_h. \quad (26)$$

Note that the OLS and WLS imputation methods described in Section 2.3 depend only on respondent data, and are therefore unaffected by multiple sources of nonresponse.

3. Application

In this section we apply the techniques developed in Section 2 to data from the Australian Agricultural and Grazing Industries Survey (AAGIS). This is an annual survey carried out by the Australian Bureau of Agricultural and Resource Economics (ABARE). The survey covers Australian farms involved in cereal crops and livestock production. AAGIS data from an earlier survey were used in a case study of single source nonresponse imputation by Chambers (1988).

The surveyed variable of interest here, and one for which nonresponse occurs, is total farm debt (Y). Each sampled farm i in the AAGIS has weight $a_i > 0$ associated with it. This weight is calculated by a modelling procedure (Bardsley and Chambers 1984) which reflects both the sampling fraction within the farm's stratum as well as its associated size variables. In forming estimates of population averages for variables (e.g., Y above) for which there is nonresponse, current practice is simply to delete sample farms for which the variable is not available, and not reweight the remaining responding farms. Mean farm debt is therefore estimated by

$$\hat{M} = \sum_{i \in s_0} a_i Y_i \Big/ \sum_{i \in s_0} a_i \quad (27)$$

where s_0 denotes those sample farms which provided debt information in the survey. It can be seen that (27) implicitly imputes the weighted mean $\hat{M}$ for debt nonrespondents.

This approach can lead to bad item imputations, as well as bad overall estimates of average farm debt, for two reasons. The first, as already noted, is that there can be a relationship between the response probability and the value of the debt variable, given the size characteristics of the farm, leading to non-ignorable nonresponse and hence a biased estimate. In this situation the debt of nonrespondents will actually follow a different regression model to that of respondents. The second is that even if no such relationship exists, nonresponse can still lead to a biased estimate because of differences between the size characteristics of respondent and nonrespondent farms. This is the so called "unconfounded" or "missing at random" situation (Pfeffermann 1988;

Rubin 1976). In either case, it should be possible to improve on (27) by modelling the relationship between farm debt and other covariates for which there is no nonresponse, including a nonresponse component to allow for possible non-ignorable nonresponse, and generating imputed values for the nonresponding farms according to the theory developed in the previous section.

In practice, debt nonresponse for AAGIS farms is allocated to one of three different categories:

i. Business structure too complex. Debt exists but not collected.

ii. Refusal by sample farm to provide debt information.

iii. Sample farm has debt but is unsure about debt details.

The distinction between these three types of nonresponse is made because there is reason to suppose that very different debt models and nonresponse mechanisms may apply to them. One would expect all three categories of nonrespondents to have nonrepresentative size characteristics so that the regression stage of the model should improve the imputation process. However, one would not expect nonresponse to be non-ignorable in all cases. In particular, category (iii) nonrespondents typically represent farms who are willing to provide debt information, but whose accounts are unavailable at the time of interview. Debt details for some of these category (iii) nonrespondents are obtained from their accountants at a later date, thus providing data for checking on the adequacy of the nonresponse imputation carried out for these farms. Since such farms represent "late" respondents, rather than nonrespondents, it is not unreasonable to expect that the selectivity bias adjustment in the procedures described in Sections 2.2 and 2.3 should not improve the imputations for category (iii) nonrespondents in these data.

The imputation formulae derived in Section 2.5 implicitly assumed that it was possible to decide on the underlying method by which nonresponse categories were allocated. In particular, it was suggested that units are allocated to these categories in a sequential manner. Such sequential behaviour is not inconsistent with the ordering of nonresponse categories above, and this approach was therefore taken. Other, more complex allocation schemes are possible to formulate for this situation, but it is expected that the difficulties in applying formulae analogous to (24) would outweigh any potential advantages.

By definition, nonrespondent farms in categories (i) and (iii) have non-zero debt. It was assumed that the category (ii) nonrespondent farms also have non-zero debt, as a farm with zero debt is unlikely to refuse to provide debt information. All responding farms with zero debt (about 20% of the sample respondents) were therefore excluded from the modelling process. This resulted in an effective sample size of 753 consisting of 626 respondents, 33 category (i), 26 category (ii), and 68 category (iii) nonrespondents.

Various combinations of predictor variables were examined. For the probit component of the model, the final set of predictor variables used in the W_i were: wheat production, number of sheep, number of beef cattle, and two zero-one flags indicating whether a farm was new to the sample and whether it was a family farm. For the regression stage, the final set of X_i predictor variables were: interest paid, total cash costs, total cash receipts, depreciation, and change in land value, based on a one-year period. Since a preliminary inspection of the regression of Y_i on X_i had indicated substantial residual heteroskedasticity, all variables

Table 2. Total debt ($1000's) of the 29 category (iii) nonrespondents for which debt details eventually became available

	Poststratification scheme		
	L1	L2	L3
Actual Debt	8,140	8,140	8,140
Current Procedure	2,330	2,330	2,330
OLS	7,640	7,030	7,200
WLS	7,000	7,320	7,260
Within Stratum	26,150	16,500	9,150
Across Stratum	6,960	6,800	7,380

included in the regression analysis (including farm debt) were transformed to the logarithmic scale before fitting (2). A square root transformation was also investigated, but did not control the heteroskedasticity as well. Since imputations are required in the original untransformed scale, all values generated by the imputation procedure applied to the transformed variables were corrected for transformation bias before use as imputations in the original scale.

The analysis was performed at three different levels of poststratification, denoted L1, L2, and L3 in Table 2 above. These correspond to increasing levels of aggregation of the sample farms on a geographic by industry basis. At each level, the analysis compared the current mean imputation procedure to the within stratum, across stratum, OLS, and WLS imputation techniques. (In equation (22), n_{h0} was replaced with $(n_{h0} - p - 1)$ to reduce the estimator's bias.)

As noted earlier, the actual data available for this analysis allowed a partial verification of these imputation techniques. For nonrespondents in category (iii), debt details are sometimes obtained at a later date, through direct access to farmers' accounts. Consequently, the actual debt values for 29 of the 68 category (iii) nonrespondents eventually became available. By comparing the results of the different imputation schemes for these late respondents, as shown in Table 2, an idea of their performance can be gained.

Any general conclusions concerning the effectiveness of the various imputation methods based on the data in Table 2 must be tempered with caution since 29 is a small number of observations and, as we shall see later, those category (iii) farms for which debt data are finally obtained are certainly not representative of all category (iii) nonrespondents. However, the following points are worth making:

- The within stratum imputation procedure is quite unstable.
- The current practice of using the mean debt of respondents to impute for the debts of nonrespondents also appears to be inferior to both the OLS and WLS approaches. (However, this observation needs to be qualified: The data used in simulating the current approach in Table 2 were based on preliminary debt values obtained from responding farms. These values are subsequently updated as accountant data become available for these farms. It is known that this updating process leads to an overall increase in debt values, so that imputations derived from the current approach can be expected to underestimate final debt values.)
- There are only marginal differences in performance between the across stratum imputation method, which allows for non-ignorable nonresponse, and the OLS and WLS imputation methods, both of which assume ignorable nonresponse. From this we conclude that the probit stage (which is meant to correct for non-ignorable nonresponse) has not improved debt imputation for

Table 3. Estimated mean debts and standard errors[1] ($1000's) based on imputed data[2]

	Current Procedure	WLS	Across Stratum
New South Wales	88.3	89.9	90.8
	(8.4)	(7.7)	(7.4)
Victoria & Tasmania	45.2	47.7	47.8
	(6.1)	(5.9)	(6.0)
Queensland	101.0	102.9	102.8
	(25.1)	(22.5)	(22.5)
South Australia	81.8	82.4	83.4
	(5.6)	(4.9)	(4.8)
Western Australia & Northern Territory	109.9	115.3	113.5
	(14.6)	(13.0)	(13.2)
Australia	80.5	83.6	83.7
	(5.4)	(5.0)	(5.0)
Type (i) nonresponse	80.5	58.4	42.1
	(5.4)	(2.0)	(10.0)
Type (ii) nonresponse	80.5	56.4	63.3
	(5.4)	(1.7)	(8.3)
Type (iii) nonresponse	80.5	156.9	159.8
	(5.4)	(4.6)	(8.1)

Note: 1. Standard errors computed via jackknifing.
2. L2 poststratification for WLS and across stratum procedures.

these 29 units. This result was not unexpected since, as mentioned earlier, there was no reason to expect category (iii) nonresponse to be non-ignorable.

- The OLS and WLS procedures give similar results for these farms.

Given the observed instability of the within stratum procedure in Table 2 and since the OLS procedure is equivalent to the within stratum procedure without the probit stage, only the current mean imputation methodology and the WLS and across stratum procedures were investigated further. In the case of the latter two imputation procedures, we also confined attention to the L2 poststratification, since this seemed to achieve the best balance between allowing different regression regimes for different farm "types" and large enough sample sizes within poststrata so as to stabilize the maximum likelihood estimates calculated under these imputation procedures.

Table 3 shows the overall mean estimates generated by these three imputation schemes using all the data after deletion of obvious outliers. The numbers in brackets are jackknifed estimates of standard error for these estimates. Note that these imputed means are all weighted as in (27) and include zero debt respondents.

The data in Table 3 clearly show the current imputation procedure leads to estimates of average debt that are biased downwards for category (iii) nonrespondents. On the other hand, category (i) and category (ii) nonrespondents appear to have smaller average debts than suggested by the current procedure. Furthermore, estimates of average debt based on across stratum imputed debt values for all types of nonresponse are less stable than those generated

by the WLS procedure. For category (ii) and (iii) nonresponse it was therefore concluded that the probit stage of the across stratum imputation approach had no significant modelling advantage over the simpler WLS procedure, that is, there appeared to be no evidence that the nonresponse was not missing at random. For category (i) nonresponse, although the standard error of the across stratum imputed mean was high ($10,000) the estimate was 1.6 standard errors away from the WLS estimate. While this result is inconclusive, it suggests that this category of nonresponse is non-ignorable (so that the probit stage is reducing bias in this case), but this is being masked by the small number of data points (33 category (i) nonrespondents) and the conservative nature of the jackknife variance estimator used to assess the significance of the bias.

As an aside, it is interesting to note that the 29 category (iii) nonrespondents for whom debt details subsequently became available, have disproportionately large debts when compared with the 68 category (iii) nonrespondents in the survey in total. This can be seen by comparing mean debt estimates for the WLS and across stratum imputations of category (iii) nonrespondents in Table 3 ($156,900 and $159,800) with the corresponding imputed average per farm debt implied by the data for the L2 stratification in Table 2 ($252,413 and $234,483). Thus although the debt data for category (iii) farms can be considered as missing at random, it is clear that there is a large variation in the X_i values for these farms which is exploited by regression imputation, but not by mean imputation.

4. Conclusions

Recently developed imputation techniques (Chambers 1988) have combined the use of linear regression adjusted for sample design effects and a normal selection mechanism for non-ignorable nonresponse. This paper has streamlined the use of the iterative EM algorithm in fitting this model as well as developed it further to account for multiple sources of nonresponse.

A case study on farm debt nonresponse imputation in the Australian Agricultural and Grazing Industries Survey has given mixed support for these techniques. It has demonstrated that regression based imputation with allowance for sample design effects can improve considerably over less sophisticated mean imputation techniques. It has also given some indication that inclusion of a selection mechanism can further reduce bias due to non-ignorable nonresponse. However, the price paid for this bias reduction is a substantial increase in variability of the survey estimates of interest, so that the overall benefit is debatable. Thus one could argue that a "missing at random" imputation strategy (e.g., WLS) should be the method of choice unless there are strong a priori reasons to believe that nonresponse is strongly non-ignorable, in which case an imputation methodology like the across stratum method investigated in this paper could be considered. Further studies may clarify this.

Finally, both an associate editor and a referee have queried the sensitivity of the suggested imputation procedures to the assumption of normality. In the context of the regression model specification, this assumption relates to the distribution of the *population* regression residuals *after* appropriate transformations of Y_i and X_i, and is therefore unremarkable. What is of some concern is the assumption of a selection mechanism for the non-ignorable nonresponse. As Little and Rubin (1987) point out, finding the "right" transformation under selection type nonresponse (that is,

one that transforms the distribution of the population residuals to normality but leaves sample selection effects unchanged) may be impossible given only respondent data since any a priori non-normality in these data will be confounded by non-normality induced by the nonresponse selection mechanism. This is certainly an issue that needs to be considered if there is extensive nonresponse. However, for the application we have in mind (the AAGI survey), the nonresponse is not extensive, and there does not seem to be any reason to worry unduly about such confounding in this case.

5. References

Bardsley, P. and Chambers, R. L. (1984). Multipurpose Estimation from Unbalanced Samples. Journal of the Royal Statistical Society, Ser. C, 33, 290–299.

Chambers, R.L. (1986). Design-Adjusted Parameter Estimation. Journal of the Royal Statistical Society, Ser. A, 149, 161–173.

Chambers, R.L. (1988). Design-Adjusted Regression With Selectivity Bias. Applied Statistics, 37, 323–334.

Chapman, D.W. (1976). A Survey of Nonresponse Imputation Procedures. In Proceedings of the Social Statistics Section, American Statistical Association, 245–251.

Chiu, H.Y. and Sedransk, J. (1986). A Bayesian Procedure for Imputing Missing Values in Sample Surveys. Journal of the American Statistical Association, 81, 667–676.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood for Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Ser. B, 39, 1–38.

Ford, B. (1983). An Overview of Hotdeck Procedures. In Incomplete Data in Sample Surveys: Theory and Bibliographies, New York: Academic Press, 185–207.

Heckman, J.J. (1979). Sample Selection Bias as a Specification Error. Econometrica, 47, 153–161.

Kotz, S. and Johnson, N.L. (1986). Quantal Response Analysis. In Encyclopedia of Statistical Sciences, New York: John Wiley.

Little, R.J.A. (1983). Superpopulation Models for Nonresponse. In Incomplete Data in Sample Surveys: Theory and Bibliographies, New York: Academic Press, Chapters 20–22.

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

Oh, H.L. and Scheuren, F.J. (1983). Weighting Adjustment for Unit Nonresponse. In Incomplete Data in Sample Surveys: Theory and Bibliographies, New York: Academic Press, 143–184.

Pfeffermann, D. (1988). The Effect of Sampling Design and Response Mechanism on Multivariate Regression-Based Predictors. Journal of the American Statistical Association, 83, 824–833.

Platek, R. and Gray, G.B. (1983). Imputation Methodology. In Incomplete Data in Sample Surveys: Theory and Bibliographies, New York: Academic Press, Chapter 17.

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

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

Received July 1989
Revised December 1990