Imputation of Missing Values When the Probability of Response Depends On the Variable Being Imputed

JOHN S. GREENLEES, WILLIAM S. REECE, and KIMBERLY D. ZIESCHANG*

A method is developed for imputing missing values when the probability of response depends upon the variable being imputed. The missing data problem is viewed as one of parameter estimation in a regression model with stochastic censoring of the dependent variable. The prediction approach to imputation is used to solve this estimation problem. Wages and salaries are imputed to nonrespondents in the Current Population Survey and the results are compared to the nonrespondents' IRS wage and salary data. The stochastic censoring approach gives improved results relative to a prediction approach that ignores the response mechanism.

KEY WORDS: Nonresponse; Imputation; Prediction approach; Censoring; Current Population Survey.

1. INTRODUCTION

There is a large literature on the problem of parameter estimation with incomplete data, but with few exceptions this literature treats the case in which the missing values are "missing at random." Rubin (1976, 1981) provides a formal probability model of the missing data problem in general and derives the conditions under which inferences can be made about the distribution of the data while the process causing the omission of data is ignored. He also points out (Rubin 1978, p. 25) that there is very little literature on parameter estimation in situations in which the mechanism causing the values to be missing is not ignorable. The purpose of this article is to treat estimation and imputation in one of the cases in which the mechanism is not ignorable—the case in which the probability of nonresponse for the variable of interest depends upon the value of that variable.

In Section 2 we discuss the problems of estimation and imputation, first in the case in which the mechanism causing the omission of data is ignorable and then in one of the cases in which it is not ignorable. Section 3 presents the derivation of the maximum likelihood estimator of the parameters of our model and shows how these can be used for imputation. In Section 4 we apply our model to the problem of imputing missing income data in the Current Population Survey (CPS). Information from a secondary source on the missing values allows a direct test of our imputation procedure. Section 5 contains a summary of our results and our conclusions.

2. ESTIMATION AND IMPUTATION WITH INCOMPLETE SAMPLE SURVEY DATA

2.1 Ignorable Response Mechanism

The literature on estimation with incomplete sample survey data is generally concerned with estimating the population total $T = \sum_{i=1}^{N} Y_i$ of some variable Y for a population of size N. With complete data, the estimation of T is to be achieved by taking a sample of size n and calculating the value of some estimator. This procedure is rendered impossible by the failure of some portion of the sample to respond to the survey question on Y_i, so that only n_r complete observations are obtained. In the remainder of this section we will assume that the simple expansion estimator, $T_1 = N \sum_{i=1}^{n} Y_i/n$, is to be used in conjunction with a procedure for handling nonresponse. There are many ways to proceed. An obvious choice is to ignore the nonresponse problem and use the n_r complete observations as the "sample:" $T_2 = N \sum_{i\in\Omega} Y_i/n_r$ where Ω is the set of indices for respondents in the sample. This estimator is equivalent to imputing the average Y for the respondents to the nonrespondents, and can often yield an unsatisfactory estimate if the nonrespondents are systematically different from the respondents on the variable of interest.

Another choice when information on auxiliary variables is available for the sample observations is to estimate T with a poststratification estimator. This approach is discussed in detail by Schaible (1979), Brewer (1979), and Oh and Scheuren (1981). Each respondent observation is weighted by the inverse of the respondent proportion of the observations in its cell or stratum, which is defined on the auxiliary variables. If there are K cells with n_{rk}

* John S. Greenlees and Kimberly D. Zieschang are economists, Office of Prices and Living Conditions, U.S. Bureau of Labor Statistics (BLS), Washington, DC 20212. William S. Reece is Chief, Economics Studies Branch, Common Carrier Bureau, Federal Communications Commission (FCC), Washington, DC 20554. Research for this article was undertaken while all three authors were members of the BLS's Division of Price and Index Number Research. The authors would like to thank Robert Gillingham, Wesley Mellow, Daniel Villegas, and two anonymous referees for helpful discussions on the research presented here and comments on an earlier draft of this paper. The views expressed are those of the authors and do not reflect the policies or views of the BLS, the FCC, or other members of their staffs.

June 1982, Volume 77, Number 378
Applications Section

respondents and n_k total sample observations in cell k, and Ω_k is the respondent portion of cell k, the poststratification estimator is

$$T_3 = N \sum_{k=1}^{K} [(n_k/n_{rk}) \sum_{i \in \Omega_k} Y_{ki}/n].$$

This estimator is equivalent to that resulting from imputing the average of the respondent observations in each cell to each nonrespondent in that cell. A similar approach is the "hot deck" approach, which consists of imputing a randomly selected cell respondent's value to each nonrespondent. In doing this, one is implicitly modeling the nonresponse mechanism by assuming that the probability of nonresponse may vary among cells but not within cells. If the probability of response varies with a continuous auxiliary variable, the poststratification approach breaks down, because the cells cannot be constructed, there being an infinite number of them.

Another approach that makes use of auxiliary data on other variables that are available for the entire sample is known as the prediction approach. (See Royall 1970; Royall and Herson 1973a,b; and Hartley and Sielken 1975.) In this approach, the analyst specifies a probability model generating the population from which the sample is taken. Usually the variable of interest is assumed to be determined by a regression on the auxiliary variables:

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i,$$

where $\mathbf{X}_i$ is the $1 \times p$ vector of values of the auxiliary variables for the ith individual and ϵ_i is a random disturbance with zero mean. The $p \times 1$ parameter vector $\boldsymbol{\beta}$ is estimated using the response sample, and the estimated parameter $\hat{\boldsymbol{\beta}}$ is used along with the auxiliary variables to predict a value of Y for each nonrespondent. Let Γ be the set of indices of the nonrespondents. The resulting estimator for the population total is

$$T_4 = N(\sum_{i \in \Omega} Y_i + \sum_{i \in \Gamma} \mathbf{X}_i\hat{\boldsymbol{\beta}})/n.$$

2.2 Nonignorable Response Mechanism

Unfortunately, all these approaches will be systematically in error if the probability of response varies with Y, a case that is very rarely considered explicitly but sometimes seems reasonable. The poststratification approach breaks down because one would need to stratify by Y, which is, of course, unavailable for the nonrespondents. The usual prediction approach breaks down for two reasons. First, the usual procedures for estimating $\boldsymbol{\beta}$ fail because the dependent variable in the regression equation is subject to stochastic censoring when the probability of response depends on Y. Second, in this case the expected value of Y_i given $\mathbf{X}_i$ for the nonrespondents is not $\mathbf{X}_i\boldsymbol{\beta}$ even given an unbiased estimator of $\boldsymbol{\beta}$. These problems result from the failure of the requirements that the disturbance, ϵ_i, be uncorrelated with $\mathbf{X}_i$ and have zero mean. When the probability of response is, for example, inversely related to Y, the expected value of the disturbance conditional on response is negative. Further, positive disturbances will be less likely to be observed with larger values of variables with positive coefficients in the regression equation than with smaller values. These results occur because individuals with positive disturbances or larger values of variables with positive regression coefficients are more likely to be nonrespondents, and hence are less likely to be in the sample from which the regression is estimated.

An example of this problem is encountered in the econometric literature on labor supply, which treats the case in which we are less likely to observe an individual's wage when he or she has a lower potential wage, since in this case the individual is less likely to be employed. This problem was identified by Gronau (1973), and solutions to this and similar estimation problems have been developed by Heckman (1974, 1976, 1979), Nelson (1977), Lee (1979), Hausman and Wise (1977), and others. Indeed, Little (1979, 1981) and Morris (1979) have mentioned this econometric literature as providing a possible solution to the estimation problem we face. However, they are not encouraging in this regard. As Little points out, "the problem with these models is that estimation is highly sensitive to unverifiable assumptions about the distribution of the underlying uncensored data . . ." (1979, p. 292). The authors mentioned earlier assume that the disturbances in the regression model are normally distributed and derive the maximum likelihood estimators of the parameters of the model. Rubin (1978) provides a simple example that illustrates the problem with the normality assumption:

> Suppose that we have a population of 1000 units, try to record a variable Z, but half of the units are nonrespondents. For the 500 respondents, the data look half-normal. Our objective is to know the mean of Z for all 1000 units. Now, if we believe that the nonrespondents are just like the respondents except for a completely random mechanism that deleted values (i.e., if we believe that mechanisms are ignorable), the mean of the respondents, that is, the mean of the half-normal distribution, is a plausible estimate of the mean for the 1000 units in the population. However, if we believe that the distribution of Z for the 1000 units in the population should look more or less normal, then a more reasonable estimate of the mean for the 1000 units would be the minimum observed value because units with Z values less than the mean refused to respond. Clearly, the data we have observed cannot distinguish between these two models except when coupled with prior assumptions. [p. 22]

Morris (1979) notes the same problem with the maximum likelihood methods used in the econometric stochastic censoring literature and insists that ". . . the user of such methods must rely on solid information from other sources about the assumed distribution . . . " (p. 463).

In this article we propose to test on a particular data set the usefulness of a stochastic censoring model, similar to those referenced earlier, for imputing when the probability of response depends on the variable being imputed. Fortunately, for the case we consider—imputing income when the probability of response depends upon income—there is a data set containing excellent data from a secondary source on the missing values from the original survey. We are referring to the matched CPS-SSA-

IRS data set, in which the data from the March 1973 Current Population Survey are matched with administrative data from the Social Security Administration (SSA) and the Internal Revenue Service (IRS). We rely for our test on IRS wage and salary data that are available for all observations in our sample, including those who failed to respond to the CPS wage and salary question.

3. ESTIMATION OF REGRESSION PARAMETERS β AND IMPUTATION OF MISSING Y VALUES WHEN THE PROBABILITY OF RESPONSE DEPENDS ON Y

3.1 Maximum Likelihood Estimation of β

As in the previous section, we assume that for the population Y_i has a regression on the vector of auxiliary variables $\mathbf{X}_i$:

$$Y_i = \mathbf{X}_i\beta + \epsilon_i, \tag{3.1}$$

where ϵ_i is a normally distributed random disturbance with zero mean and constant variance σ^2 and is uncorrelated with ϵ_j for $i \neq j$. We also assume that the probability of response to the survey question on Y depends on Y. One way in which this can be modeled is to assume that the probability of response is a logistic function of Y and other variables Z:

$$P(R_i = 1 \mid Y_i, \mathbf{Z}_i) = \frac{1}{1 + \exp(-\alpha - \gamma Y_i - \mathbf{Z}_i\delta)}, \tag{3.2}$$

where R_i is a variable equaling unity if individual i is a respondent and zero if a nonrespondent; $\mathbf{Z}_i$ is a $1 \times m$ vector of characteristics of individual i; α and γ are scalar parameters; and δ is an $m \times 1$ parameter vector. If γ is positive, the probability of response varies directly with Y, and if γ is negative the probability of response varies inversely with Y. Note that if we had used a probit instead of logit function our model would have been essentially identical to that of Nelson (1977).

Cassel, Sarndal, and Wretman (1979) suggest a similar model of nonresponse, explicitly considering the case in which the probability of response depends on the variable being imputed. However, they propose estimating the probability of response as a function of $\mathbf{X}$ alone, rather than of $\mathbf{X}$ and Y. The overall approach taken in their article is a robust prediction approach based on estimating β using weighted least squares. However, when the probability of response depends on Y itself rather than only its systematic part (i.e., that part explained by $\mathbf{X}$), any least squares approach will yield inconsistent estimates of β because $\mathbf{X}$ and ϵ are correlated. A different estimation criterion is needed, and we have adopted the maximum likelihood criterion, assuming the disturbances are normally distributed. Our modifications of the Cassel, Sarndal, and Wretman model allow the possibility of consistent estimation of β when the probability of response depends on Y.

We assume (3.1) and (3.2) hold for all n units in the sample. However, we observe Y_i only for the n_r respondents, while we observe $\mathbf{X}_i$ for all units in the sample. Without loss of generality we can order the data so that observations 1 through n_r are the respondents. The likelihood function for this sample will then consist of the product of $n - n_r$ factors for the nonrespondents and n_r factors for the respondents. The factor of the likelihood for each of the respondents will be the product of the probability of response given income and the value of the density of income at that level:

$$L_i = \frac{1}{1 + \exp(-\alpha - \gamma Y_i - \mathbf{Z}_i\delta)} \times \frac{1}{\sigma}\phi\left(\frac{Y_i - \mathbf{X}_i\beta}{\sigma}\right), \quad i = 1, \ldots, n_r. \tag{3.3}$$

For each nonrespondent the factor of the likelihood is simply the marginal probability of nonresponse:

$$L_i = \int_{-\infty}^{\infty}\left(1 - \frac{1}{1 + \exp(-\alpha - \gamma Y - \mathbf{Z}_i\delta)}\right) \times \frac{1}{\sigma}\phi\left(\frac{Y - \mathbf{X}_i\beta}{\sigma}\right)dY, \quad i = n_r + 1, \ldots, n. \tag{3.4}$$

The likelihood function of the entire sample is thus $\prod_{i=1}^{n} L_i$, where L_i is given by (3.3) or (3.4), according to the value of i.

Maximum likelihood estimates of the parameters of this model can be found by numerically maximizing the log of this function with respect to α, γ, δ, β, and σ, given Y_i for $i = 1, \ldots, n_r$ and $\mathbf{Z}_i$ and $\mathbf{X}_i$ for $i = 1, \ldots, n$. Here we use the generalized Gauss-Newton algorithm described by Berndt, Hall, Hall, and Hausman (1974). See Appendix A for computational details.

3.2 Imputation of Y for Nonrespondents

Given the estimated parameters of the model, we can impute individual nonrespondents' Y values by assigning the mean of the distribution of Y conditional on nonresponse, the values of $\mathbf{Z}$ and $\mathbf{X}$ for that individual, and the maximum likelihood parameter estimates $\hat{\alpha}$, $\hat{\beta}$, $\hat{\gamma}$, $\hat{\delta}$, and $\hat{\sigma}$. This mean can be calculated in a straightforward way using numerical integration:

$$E(Y_i \mid \mathbf{X}_i, \mathbf{Z}_i, R_i = 0) = \frac{\int_{-\infty}^{\infty} Y\left(1 - \frac{1}{1 + \exp(-\hat{\alpha} - \hat{\gamma}Y - \mathbf{Z}_i\hat{\delta})}\right)}{\int_{-\infty}^{\infty}\left(1 - \frac{1}{1 + \exp(-\hat{\alpha} - \hat{\gamma}Y - \mathbf{Z}_i\hat{\delta})}\right)} \times \frac{\frac{1}{\hat{\sigma}}\phi\left(\frac{Y - \mathbf{X}_i\hat{\beta}}{\hat{\sigma}}\right)dY}{\frac{1}{\hat{\sigma}}\phi\left(\frac{Y - \mathbf{X}_i\hat{\beta}}{\hat{\sigma}}\right)dY}. \tag{3.5}$$

This procedure is suitable for the purpose at hand, that of evaluating the imputation bias that results from ignoring the effect of income on nonresponse. Other imputation methods will be appropriate in other contexts. In the typical situation in which the imputed values are to be

used as inputs to further statistical analyses it will be important to avoid understating the variance of Y, as would occur if the conditional expectation were imputed to each nonrespondent. To avoid this problem, missing values should be assigned draws from the distribution of income conditional on nonresponse and on the values of $\mathbf{X}$ and $\mathbf{Z}$. Using our parameter estimates, this can be accomplished as follows:

1. Draw ϵ_i from a $N(0, 1)$ generator;
2. Calculate $Y_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \hat{\sigma}\epsilon_i$ and the probability of nonresponse $P(R = 0 \mid Y_i, \mathbf{Z}_i) = 1 - 1/[1 + \exp(-\hat{\alpha} - \hat{\gamma}Y_i - \mathbf{Z}_i\hat{\boldsymbol{\delta}})]$;
3. Draw a random variable η from a uniform generator over the [0, 1] interval;
4. Keep Y_i as the imputed value for observation i if $P(R = 0 \mid Y_i, \mathbf{Z}_i) \geq \eta$, otherwise return to step 1.

While we have emphasized consistency of imputation conditional on our model, still other considerations should govern the choice of an imputation method in practice. For example, Rubin (1978) has suggested imputing multiple values for each nonrespondent, generating multiple data sets on which the users can evaluate the sensitivity of their analyses to the alternative imputed values and calculate variances of estimates that reflect the loss in precision due to missing data. This loss in precision is ignored by single imputation procedures. See Herzog and Rubin (1981) for an example of this procedure applied to CPS data.

4. IMPUTING INCOME TO NONRESPONDENTS IN THE CURRENT POPULATION SURVEY

In this section we conduct a test of the usefulness of our model for imputing incomes to actual nonrespondents. However, before doing this we conduct tests of two assumptions underlying the model. After discussing the data in the first subsection, we test whether the probability of response depends on the level of income as we have hypothesized. We then conduct an approximate test of whether the disturbances ϵ in (3.1) are normally distributed. These two tests, which require data on the nonrespondents' incomes as well as on the respondents' incomes, can be conducted because of the special nature of our data set. Then, in subsections 4.4 and 4.5, we discard the nonrespondents' income data and impute these missing values using the prediction approach, first ignoring the response mechanism and then incorporating the response mechanism by means of the method presented in Section 3. We then compare these imputed values to the actual values.

4.1 Data

The Current Population Survey is a very large multipurpose monthly survey conducted by the Census Bureau to provide data on income and employment and other characteristics of the noninstitutional U.S. population. Each year the March survey includes additional detail on income and employment for the previous calendar year. The Census Bureau and the SSA, with the aid of the IRS, have matched the March 1973 survey with data from Social Security benefit and earnings records and from federal income tax records. This is an exact match by social security number. This exact match file provides data on income in 1972 for a large sample of individuals from the CPS, where income nonresponse occurs, and from federal income tax returns, where there is complete response on income. Using this file we can construct a data set that contains income data from actual CPS income nonrespondents. The public-use file that we use is described in Aziz, Kilss, and Scheuren (1978).

We use the household head's response status on the CPS wage and salary question to indicate which households were nonrespondents. Throughout our analysis, however, our measure of income, LOGWAGE, is the logarithm of the IRS wage and salary variable, which is available for both respondents and nonrespondents. Thus, we use a single definition of the income variable for all observations. Furthermore, although we do not use the nonrespondents' wage and salary data in estimating our model parameters, the true "missing values" are available for comparison with our imputations.

If our primary purpose were to impute CPS wage and salary income to nonrespondents, we would naturally use CPS income rather than IRS income to estimate our income and response function parameters. However, our goal is to construct an experimental situation in which the usefulness of our model can be empirically evaluated. For this purpose we define a sample of IRS wage and salary income nonrespondents, not in an arbitrary way, but according to nonresponse on a related variable, CPS wage and salary income, thereby obtaining a realistic pattern of missing data. It is not necessary that the two income measures be identical. We require instead only that the probability of response on the CPS be functionally related to IRS income. This relationship is estimated in Section 4.2, which follows.

To reduce our computational burden, we first reduce the sample in size and second make it less heterogeneous to reduce the number of parameters to be estimated. From the subset of the file for which exact matches to the IRS data were actually made, we select heads of basic primary families in which the head was at least 14 years old, was married with spouse present, had a nonfarm residence, and had no farm or self-employment income. Further, the head must have been employed full time for the full year 1972 in the private nonagricultural sector and must have filed a joint tax return. Because the IRS wage and salary data include spouses' wages with the household heads' wages we select only heads of households with spouses who did not work in 1972. Finally, a few returns have unbelievably low reported IRS wage and salary figures for employees working full year and full time. We discard six observations having IRS wage

Table 1. Estimated Response Function Parameters and Standard Errors

Independent variables	Simple Logit Equation Estimated Using Both Respondent and Nonrespondent LOGWAGE Data		Response Portion of the Stochastic Censoring Model Estimated Using Only Respondent LOGWAGE Data	
	Response 1[a]	Response 2[b]	Response 1[a]	Response 2[b]
LOGWAGE	−.4032 (.1028)	−.4259 (.1195)	−.1955 (.2665)	−.4301 (.3182)
PERSONAL	.3553 (.0928)	.4038 (.1078)	.3718 (.0925)	.4080 (.1071)
AGE	−.0388 (.0040)	−.0456 (.0046)	−.0398 (.0041)	−.0456 (.0049)
EDUCATION	−.0632 (.0179)	−.0669 (.0208)	−.0802 (.0268)	−.0659 (.0318)
WHITE	.1826 (.2299)	−.4353 (.3497)	.1495 (.2349)	−.4339 (.3543)
NORTH	.2401 (.1146)	.1825 (.1302)	.2386 (.1147)	.1855 (.1304)
SOUTH	.2877 (.1235)	.3993 (.1476)	.3092 (.1256)	.4040 (.1483)
WEST	.3963 (.1442)	.3278 (.1630)	.4017 (.1444)	.3299 (.1624)
CONSTANT	7.9546 (.9204)	9.4318 (1.1022)	6.2494 (2.1947)	9.4583 (2.6659)
Sample Size	5515	5364	5515	5364

[a] Response 1 equals unity if the household head responded to the CPS wage and salary questions; zero otherwise.
[b] Response 2 equals unity if the household head responded to the CPS wage and salary questions; zero if the household head *refused* to answer.

and salary figures below $500.[1] These restrictions leave a sample of 5,515 observations.

The CPS survey procedures allow for different kinds of nonresponse. In addition to the cases response and refusal, there are cases in which the question may be unanswered for other reasons. Rather than attempt to estimate the parameters of a polytomous response function, we define two alternative response status variables, one in which any nonresponse is allowed and the second in which only refusals are counted as nonrespondents. There are 410 CPS wage and salary refusals in our sample and 151 cases in which the question is unanswered for other reasons. When the refusal variable is used, these 151 observations are deleted from the sample, leaving a sample size of 5,364.

An unfortunate feature of the data that creates additional complications at almost every stage of our analysis is the censoring of all dollar amounts at 50,000. That is, all IRS wage and salary figures exceeding $50,000 were recoded to that figure in creating the public-use tape. As a result, we must add an additional type of term to the log of the likelihood function, and modify our imputation procedures. These modifications are explained in Appendix A.

[1] A referee suggests that these anomalous observations may be due to faulty matching or prior imputation of the independent variables. In this article we have ignored these aspects of the CPS data base. Also we do not deal with sample weighting, rotation group effects, and other issues that deserve to be addressed in future research on nonresponse in large surveys with complex sample designs.

4.2 Does the Probability of Response Depend On Income?

Before estimating the parameters of the model specified in Section 3, we test the hypothesis that the probability of response depends on income. If this hypothesis is rejected, then nonresponse is plausibly ignorable and the straightforward prediction approach can be used. For this test we use the entire sample of observations, including income data from both nonrespondents and respondents. Specifically, we hypothesize that the probability of response on the CPS household head wage and salary question is a function of IRS wage and salary income as well as other variables chosen on the basis of a priori notions of the factors that might influence response behavior. The form of the relationship is assumed to be logistic, as specified in (3.2). The definitions of the two response status variables and the independent variables are given in Appendix B.

The parameters governing the probability of response for the two alternative definitions of nonresponse were estimated by maximum likelihood using a Newton-Raphson algorithm. The estimates are shown in the first two columns of Table 1. We see that for both definitions of response, the results are qualitatively the same. Most important, the LOGWAGE coefficient is negative and almost four times as large as its standard error. This evidence strongly supports the hypothesis that the probability of response depends on the wage and salary level, and indicates that the individuals with higher wages and

salaries have smaller probabilities of response. The results also indicate that the individuals interviewed in person are more likely to respond than those interviewed by telephone, that the older individuals are less likely to respond than the younger, and that those with more years of education are less likely to respond than those with fewer years of education. The results indicate that individuals in the Eastern region are less likely to respond than individuals in the other three regions and that race does not appear as a significant factor in determining the probability of response.

4.3 Are the Income Function Disturbances Normally Distributed?

As stated in Section 3.1, we assume the disturbances ϵ in the semi-logarithmic income function (3.1) are normally distributed. Methods such as ours have been criticized in the past due to their reliance on distributional assumptions that are usually, in practice, untestable. However, in our case we have auxiliary data that allow examination of distributional assumptions. In this section we estimate the parameters of the income function using the entire data set, including respondents' and nonrespondents' incomes, and examine the residuals of the equation for evidence on the distribution of ϵ.

There is a large literature in economics on estimating earnings equations. See, for example, Mincer (1974). We estimate the parameters of an equation similar to those typically specified. Our dependent variable is LOGWAGE as defined in Section 4.1, and the sample is the respondent portion of the sample described in Section 4.1. The independent variables are defined in Appendix B.

Our estimates of the parameters of the wage and salary determination equation are shown in the first column of Table 2. The parameters are estimated by maximum likelihood under the assumption of normality of the disturbances, but taking account of the censoring of income at \$50,000. See Appendix A for details. Our estimates are similar to those obtained in other studies of income determination. See, for example, Ashenfelter (1978). The parameters indicate that the effect of additional education on LOGWAGE is positive and increases with additional education, while the effect of additional years of experience is positive and decreases with additional experience. The wages of whites are significantly higher than the wages of nonwhites, and workers in the South earn significantly less than workers in the other regions. Workers residing in the suburbs of Standard Metropolitan Statistical Areas (SMSA's) earn more than workers residing in central cities, and both earn more than non-SMSA residents. The industry and occupation effects are also quite reasonable.

Table 3 compares selected percentiles of the empirical distribution function of the standardized residuals, $(Y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})/\hat{\sigma}$, from this equation, to percentiles of the standard normal distribution function. The empirical distribution function has mean near zero ($-.004$) and is approximately symmetric over most of its observed range. The symmetry result is encouraging; as noted by Rubin (1979) and Little (1979) in other contexts, an unwarranted assumption of symmetry can produce spurious evidence of relationships between the variables under study. The kurtosis of the empirical distribution is greater than that of the standard normal, due in part to the presence of some large negative residuals. A Kolmogorov test would lead to rejection at a high level of significance of the hypothesis that the observed standardized residuals are realizations of a standard normal variate. Because the residuals are censored and have unequal variances this result does not constitute a strictly valid test. But, if we take it as suggestive, the result is unfavorable to the normality assumption.

Nevertheless, some assumption concerning the distribution of ϵ must be made if we are going to assume the response mechanism is not ignorable. In practice one would not be able to test hypotheses concerning the distribution of ϵ, not having the nonrespondents' observations on the dependent variable, Y. Since normality of ϵ would most often be assumed, we maintain that assumption in the following sections, where we conduct a more direct test of the usefulness of the model—a test of its ability to impute.

4.4 Application of the Prediction Approach Under the Assumption That the Response Mechanism Is Ignorable

If the dependence of the probability of response on the level of income could be ignored, it would be appropriate to apply the prediction approach, estimating the vector of income determination parameters using only the sample of respondents. To evaluate the use of this procedure in imputing wage and salary levels to CPS wage and salary nonrespondents, we again use the IRS wage and salary variable as a proxy for the CPS variable and use the sample of 4,954 CPS respondents to estimate an income equation. The specification and estimation method are the same as those used in Section 4.3.

The results of this estimation are shown in the second column of Table 2. As might be expected, given the similarity of the samples, the parameter estimates are almost identical to those in Column 1. We use the estimated $\hat{\boldsymbol{\beta}}$ and the nonrespondents' values of the independent variables to impute the expected value of LOGWAGE for each nonrespondent. This expected value will be slightly less than $\hat{Y}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}}$ because of the censoring of LOGWAGE at log(50000); the precise imputation formula is presented in Appendix A.

The results of Section 4.2 imply that this prediction method will systematically impute income levels below the true values. The first two rows of Table 4 display the results of such a comparison. For the sample of 561 nonrespondents, the logarithm of IRS wage and salary income exceeds the imputed value by an average of .0768—

Table 2. Estimated Wage Determination Equation Parameters and Standard Errors

	Estimation Method and Sample			
	Regression		Stochastic Censoring	
Independent variables	Respondents and All Nonrespondents	Respondents	Respondents and All Nonrespondents	Respondents and Refusals
EDUCATION	−.0176 (.0095)	−.0184 (.0098)	−.0181 (.0098)	−.0178 (.0098)
EDUCATION 2	.3333 (.0422)	.3344 (.0415)	.3344 (.0416)	.3340 (.0416)
EXPERIENCE	.0344 (.0018)	.0340 (.0016)	.0342 (.0017)	.0343 (.0017)
EXPERIENCE 2	−.0551 (.0033)	−.0557 (.0030)	−.0558 (.0030)	−.0558 (.0030)
WHITE	.1566 (.0275)	.1653 (.0243)	.1650 (.0243)	.1665 (.0244)
CENTRAL CITY	.1145 (.0161)	.1126 (.0162)	.1125 (.0162)	.1129 (.0162)
SUBURB	.1792 (.0143)	.1787 (.0148)	.1796 (.0148)	.1807 (.0148)
NORTH	.0298 (.0147)	.0342 (.0160)	.0336 (.0160)	.0335 (.0160)
SOUTH	−.0801 (.0153)	−.0670 (.0160)	−.0678 (.0160)	−.0687 (.0161)
WEST	−.0062 (.0175)	.0002 (.0187)	−.0008 (.0188)	−.0012 (.0187)
PROFESSIONAL	.4032 (.0360)	.3667 (.0397)	.3661 (.0397)	.3675 (.0397)
SALES	.2215 (.0372)	.1940 (.0411)	.1947 (.0411)	.1976 (.0411)
CRAFT	.1778 (.0346)	.1576 (.0393)	.1566 (.0393)	.1576 (.0393)
LABORER	.0761 (.0433)	.0592 (.0504)	.0597 (.0504)	.0601 (.0505)
CONSTRUCTION	.2493 (.0243)	.2611 (.0226)	.2610 (.0226)	.2614 (.0226)
MANUFACTURING	.1390 (.0190)	.1462 (.0184)	.1457 (.0184)	.1461 (.0185)
TRANSPORTATION	.2143 (.0233)	.2256 (.0229)	.2247 (.0229)	.2249 (.0229)
TRADE	.0719 (.0209)	.0672 (.0186)	.0667 (.0186)	.0671 (.0186)
SERVICE	.0030 (.0554)	−.0475 (.0464)	−.0464 (.0464)	−.0465 (.0464)
CONSTANT	8.0744 (.0702)	8.0958 (.0749)	8.0936 (.0750)	8.0865 (.0752)
σ	.4108	.4003	.4004	.4007
Sample size	5515	4954	5515	5364

that is, the average underestimate of income is approximately 8 percent. The sample variance of the error is .2307, indicating that large imputation errors are present.

When we divide the mean imputation error by its standard deviation, we obtain .0768/.0203 = 3.78. Under the null hypothesis of unbiased imputation this ratio follows an asymptotic standard normal distribution. An alternative, nonparametric test uses the fact that the prediction approach with ignorable mechanism underestimates LOGWAGE in 323 of 561 cases. Under the null hypothesis that the probability of underestimate is .5, we obtain a test statistic of 3.55, which is also approximately unit normal by the normal approximation to the binomial distribution. Both of these tests lead us to reject the null

Table 3. Values of Theoretical and Empirical Distribution Functions of Standardized Income Equation Residuals

Percentile	Normal Distribution	Observed Distribution
1	−2.33	−3.15
10	−1.28	−1.03
20	−.84	−.65
30	−.52	−.39
40	−.25	−.18
50	.00	.01
60	.25	.22
70	.52	.44
80	.84	.68
90	1.28	1.10
99	2.33	2.39

hypothesis at a high level of significance, supporting our expectation that the response mechanism is not ignorable in this case.

Similar conclusions are reached using the second definition of nonresponse. The same wage equation estimates are used, since the respondent sample is unchanged. The imputation results for the refusals only are shown in the second row of Table 4. The average underestimate of income is just under 7 percent. The ratio of this mean to its standard deviation is 3.09. Income is underestimated in 236 of 410 cases, yielding an alternative, unit normal test statistic of 3.01. Again, both tests lead to rejection of the null hypothesis of unbiased imputation.

It is interesting to note that the results of this section are consistent with the experience of the Census Bureau in imputing CPS wage and salary income to survey nonrespondents. As reported by Herriot and Spiers (1975), the ratio of mean CPS imputed wages and salaries to IRS wages and salaries for CPS nonrespondents was .91, compared to .98 for respondents, which indicates a downward imputation bias of approximately seven percent.

4.5 Application of the Prediction Approach With Stochastic Censoring

In this section we present the results of applying the stochastic censoring model discussed in Section 3.1 to the simultaneous estimation of the parameters of the income and response functions using the respondents' LOGWAGE data and data on the independent variables for the respondents and nonrespondents. We then impute expected LOGWAGE for each nonrespondent and compare these imputations both to the actual values and to the values imputed in the previous section. Both definitions of nonresponse are used in turn.

Table 4. Means and Variances of Differences between Actual and Imputed LOGWAGE

Method	Mean Imputation Error	Variance of Imputation Error	t Statistic	Numbers of Cases
Prediction with ignorable mechanism				
RESPONSE 1	.0768	.2307	3.78	561
RESPONSE 2	.0687	.2015	3.09	410
Prediction with stochastic censoring				
RESPONSE 1	.0455	.2308	2.24	561
RESPONSE 2	.0001	.2015	0.00	410

The specifications of the income and response functions are the same as discussed earlier. The estimated income equation parameters are shown in the third and fourth columns of Table 2, and the estimated response function parameters are shown in the third and fourth columns of Table 1. The estimated parameters of the income equations are almost identical to the estimates obtained using the respondent sample only. The estimated response functions are also similar to those obtained previously with the exception of the LOGWAGE variable. Although the coefficient on LOGWAGE in the response function for all nonrespondents (third column of Table 1) has the correct sign, it is only about one-half as large as the estimate in the first column, and the asymptotic standard error is much larger. The results obtained using the second definition of nonresponse are more satisfactory, as might be expected given the clearer behavioral dichotomy between respondents and refusals. The LOGWAGE coefficients in the response functions for refusals only (the second and fourth columns of Table 1) are nearly equal, although the asymptotic standard error is much larger when only respondent income data is used in the estimation process.

The negative signs on the LOGWAGE coefficients obtained using our stochastic censoring model indicate that relative to respondents the nonrespondents are more likely to have algebraically larger disturbances in the income equation at each level of $\mathbf{X}\boldsymbol{\beta}$. As a result, we will be imputing higher incomes to the nonrespondents using (3.5) than we did in Section 4.4. Since the earlier imputations were biased downward, we expect to do better.

The formula used in imputing LOGWAGE under stochastic censoring is given in Appendix A and is based on (3.5). The means and variances of our imputation errors are displayed in the last two rows of Table 4. For the first nonresponse definition, imputed income in each case is between .029 and .034 higher than before. As a result, the mean error of imputation using stochastic censoring is 41 percent smaller than the mean error achieved assuming an ignorable response mechanism, and the variances of the sample error distributions are nearly identical. The stochastic censoring imputations still appear to have a significant downward bias, due to the relatively low estimated coefficient on LOGWAGE in the response function. Row 4 of Table 4 displays the imputation summary for the refusal sample. The bias is nearly eliminated, with the mean error of imputation being only .0001 as compared to .0687 using the model of Section 4.4. This striking improvement results from the accuracy with which we were able to estimate the effect of LOGWAGE on the probability of refusal.

5. SUMMARY AND CONCLUSIONS

Our most important results can be summarized as follows:

We develop a method for imputing missing values when the probability of nonresponse depends on the variable being imputed. In this case, since we can develop (actually, borrow) a model of the determination of the variable of interest, income, we can view the missing data problem as one of estimation with stochastic censoring of the dependent variable. Having solved this estimation problem, we can use the prediction approach to impute the logarithms of missing income values.

The existence of a data set with reliable auxiliary information on missing income values allows us to test some of our assumptions as well as test the usefulness of our imputation procedure. We test the hypothesis that the probability of nonresponse to the income question depends on the level of income, and find a strong tendency for those with higher incomes to respond less frequently. While the implications of the distributional assumptions underlying our stochastic censoring estimation procedure are complex and are not formally tested, informal examination of the data suggests that the distribution of the residuals of the semi-logarithmic income determination equation is roughly symmetric, but not normal. Nevertheless, application of our procedures to the imputation of missing income values yields significantly better imputations than a prediction approach that ignores the mechanism causing the nonresponse.

We conclude that the stochastic censoring approach to imputing missing values has potential for improving on commonly used imputation procedures that rely on the assumption that the probability of nonresponse does not depend on the variable being imputed. Application of this method requires a model of the determination of the variable of interest and a model of the nonresponse mechanism. Further, except in special cases such as the one we consider, the researcher will have little or no opportunity to evaluate either the validity of his distributional assumptions or the accuracy of his imputations. However, the results presented here provide empirical support for continued research using models such as ours, particularly research on the robustness of parameter estimates and imputed values. Specifically, the work of Rubin (1978) and Herzog and Rubin (1981) in multiple imputation provides an avenue for comparing the variability of the imputations of competing models.

APPENDIX A: LIKELIHOOD FUNCTIONS AND IMPUTATION FORMULAS WITH TOPCODED DEPENDENT VARIABLE

Our statistical analysis of the matched CPS-SSA-IRS data base is complicated by the "topcoding" of all income variables. That is, values exceeding $50,000 have been coded as $50,000, and as a result the variable LOGWAGE is censored from above at approximately 10.82 for a small number of individuals (47 respondents, 8 refusals, and 8 other nonrespondents in the wage and salary sample). This censoring is in addition to the stochastic censoring due to nonresponse which is the primary subject of this article.

The topcoding of LOGWAGE is ignored in the logit analysis reported in Section 4.2. The coefficients in the first two columns of Table 1 can therefore be expected to contain some small degree of error. Censoring of the dependent variable is, however, taken into account explicitly in the analyses of Sections 4.3 and 4.4. The income equation is estimated by the maximum likelihood approach, using the entire sample in Section 4.3 and the respondent sample in Section 4.4. The likelihood of the ith observation under this specification is given by

$$L_i = \begin{cases} \frac{1}{\sigma}\phi\left(\frac{Y_i - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right), & \text{for } Y_i < \log(50000) \\ \int_{\log(50000)}^{\infty} \frac{1}{\sigma}\phi\left(\frac{Y - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right) dY, & \text{for } Y_i = \log(50000) \end{cases} \tag{A.1}$$

This is the well-known "Tobit" model developed by Tobin (1958). For the nontopcoded observations, the likelihood value is identical to that of the standard linear regression model, and, in fact, the estimates obtained from the specification (A.1) are close to ordinary least squares estimates for our data because of the small number of topcoded observations.

In Section 3.1 the likelihood of a respondent observation in the fully specified stochastic censoring model is given as

$$L_i = \frac{1}{1 + \exp[-\alpha - \gamma Y_i - \mathbf{Z}_i\boldsymbol{\delta}]} \frac{1}{\sigma}\phi\left(\frac{Y_i - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right). \tag{A.2}$$

This again must be modified to take account of topcoding, and the formula used to derive the estimates reported in Section 4.5 is

$$L_i = \begin{cases} \frac{1}{1 + \exp(-\alpha - \gamma Y_i - \mathbf{Z}_i\boldsymbol{\delta})} \times \frac{1}{\sigma}\phi\left(\frac{Y_i - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right), & \text{for } Y_i < \log(50000) \\ \int_{\log(50000)}^{\infty} \frac{1}{1 + \exp(-\alpha - \gamma Y - \mathbf{Z}_i\boldsymbol{\delta})} \times \frac{1}{\sigma}\phi\left(\frac{Y - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right) dY, & \text{for } Y_i = \log(50000). \end{cases} \tag{A.3}$$

Topcoding has no effect on the likelihood of a nonresponse observation. However, the censoring of incomes above $50,000 does become important when we employ our parameter estimates to impute the incomes of nonrespondents and compare these imputations to the actual

(sometimes topcoded) values. To minimize the expected squared error of imputation, we calculate the mean of the distribution of the logarithm of income given nonresponse and the presence of topcoding. The expression for this mean is a modification of (3.5):

$$
\begin{aligned}
&E(Y_i \mid \mathbf{X}_i, \mathbf{Z}_i, R_i = 0) \\
&= \left\{ \int_{-\infty}^{\log(50000)} Y\left(1 - \frac{1}{1 + \exp(-\alpha - \gamma Y - \mathbf{Z}_i\boldsymbol{\delta})}\right) \right. \\
&\quad \times \frac{1}{\sigma}\phi\left(\frac{Y - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right) dY + \log(50000) \\
&\quad \times \int_{\log(50000)}^{\infty} \left(1 - \frac{1}{1 + \exp(-\alpha - \gamma Y - \mathbf{Z}_i\boldsymbol{\delta})}\right) \\
&\quad \left. \times \frac{1}{\sigma}\phi\left(\frac{Y - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right) dY \right\} \\
&\quad \div \int_{-\infty}^{\infty} \left(1 - \frac{1}{1 + \exp(-\alpha - \gamma Y - \mathbf{Z}_i\boldsymbol{\delta})}\right) \\
&\quad \times \frac{1}{\sigma}\phi\left(\frac{Y - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right) dY.
\end{aligned}
\tag{A.4}
$$

Under the model of Section 4.4, in which the response mechanism is ignorable, the formula for imputation is identical to (A.4) except that the logistic probability-of-response terms are deleted. The formula then simplifies to

$$
\begin{aligned}
E(Y_i \mid \mathbf{X}_i) &= \int_{-\infty}^{\log(50000)} Y \frac{1}{\sigma}\phi\left(\frac{Y - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right) dY \\
&\quad + \log(50000) \int_{\log(50000)}^{\infty} \frac{1}{\sigma}\phi\left(\frac{Y - \mathbf{X}_i\boldsymbol{\beta}}{\sigma}\right) dY.
\end{aligned}
\tag{A.5}
$$

Using (A.4) or (A.5) the imputed income is always less than log(50000).

In our estimation program, three-point Gaussian quadrature is used to approximate all definite integrals with finite upper or lower bounds. Integrals with infinite upper and lower bounds were evaluated using a Hermite polynomial approximation.

APPENDIX B: DEFINITIONS OF VARIABLES

RESPONSE 1:	Unity if the household head responded to the CPS wage and salary questions; zero otherwise.
RESPONSE 2:	Unity if the household head responded to the CPS wage and salary questions; zero if the household head *refused* to answer.
LOGWAGE:	The log of the IRS wage and salary variable.
PERSONAL:	Unity if the March CPS interview was a personal interview; zero otherwise.
AGE:	The age of the household head.
WHITE:	Unity if the race of the household head is white; zero otherwise.
NORTH:	Unity if the household resides in the North Central region; zero otherwise.
SOUTH:	Unity if the household resides in the South region; zero otherwise.
WEST:	Unity if the household resides in the West region; zero otherwise.
EDUCATION:	The number of years of education completed by the household head.
EDUCATION 2:	EDUCATION squared ($\times 10^{-2}$).
EXPERIENCE:	AGE − EDUCATION − 6. (intended to represent years of experience in the labor market).
EXPERIENCE 2:	EXPERIENCE squared ($\times 10^{-2}$).
CENTRAL CITY:	Unity if the household resides in the central city of an SMSA; zero otherwise.
SUBURB:	Unity if the household resides in the ring of the SMSA; zero otherwise.
PROFESSIONAL:	Unity if the household head's occupation is professional or managerial; zero otherwise.
SALES:	Unity if the household head's occupation is sales or clerical; zero otherwise.
CRAFT:	Unity if the household head's occupation is craft or operative; zero otherwise.
LABORER:	Unity if the household head's occupation is laborer; zero otherwise.
CONSTRUCTION:	Unity if the household head is employed in the construction or mining industries; zero otherwise.
MANUFACTURING:	Unity if the household head is employed in the manufacturing industry; zero otherwise.
TRANSPORTATION:	Unity if the household head is employed in the transportation, communication, or utilities industries; zero otherwise.
TRADE:	Unity if the household head is employed in the wholesale or retail trade industries; zero otherwise.
SERVICE:	Unity if the household head is employed in the personal serv-

ice, entertainment, or recreation service industries; zero otherwise.

CONSTANT: Unity for all observations.

[*Received August 1980. Revised June 1981.*]

REFERENCES

ASHENFELTER, ORLEY (1978), "Union Relative Wage Effects: New Evidence and a Survey of Their implications for Wage Inflation," in *Econometric Contributions to Public Policy*, eds. R. Stone and W. Peterson, New York: MacMillan.

AZIZ, FAYE, KILSS, BETH, and SCHEUREN, FREDERICK (1978), *1973 Current Population Survey—Administrative Record Exact Match File Codebook, Part I—Code Counts and Item Definitions*, Washington, D.C.: U.S. Department of Health, Education and Welfare.

BERNDT, ERNST K., HALL, BRONWYN H., HALL, ROBERT E., and HAUSMAN, JERRY A. (1974), "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement*, 4, 653–665.

BREWER, KENNETH R. (1979), "Discussion," in *Symposium on Incomplete Data: Preliminary Proceedings*, Washington, D.C.: U.S. Department of Health, Education, and Welfare, 219–224.

CASSEL, CLAES-MAGNUS, SARNDAL, CARL-ERIK, and WRETMAN, JAN H. (1979), "Some Uses of Statistical Models in Connection with the Nonresponse Problem," in *Symposium on Incomplete Data: Preliminary Proceedings*, Washington, D.C.: U.S. Department of Health, Education, and Welfare, 188–215.

GRONAU, REUBEN (1973), "The Effect of Children on the Housewife's Value of Time," *Journal of Political Economy*, 81, Supplement, 168–199.

HARTLEY, H.O., and SIELKEN, R.L. (1975), "A 'Super-population Viewpoint' for Finite Population Sampling," *Biometrics*, 31, 411–422.

HAUSMAN, JERRY A., and WISE, DAVID A. (1977), "Social Experimentation, Truncated Distributions, and Efficient Estimation," *Econometrica*, 45, 919–938.

HECKMAN, JAMES (1974), "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, 42, 679–694.

——— (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.

——— (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

HERRIOT, R.A., and SPIERS, E.F. (1975), "Measuring the Impact on Income Statistics of Reporting Differences between the Current Population Survey and Administrative Sources," *Proceedings*, American Statistical Association Social Statistics Section, 147–158.

HERZOG, THOMAS N., and RUBIN, DONALD B. (1981), "Using Multiple Imputations to Handle Nonresponse in Sample Surveys," paper for Panel on Incomplete Data, Committee on National Statistics, National Academy of Sciences.

LEE, LUNG-FEI (1979), "Identification and Estimation in Binary Choice Models with Limited (Censored) Dependent Variables," *Econometrica*, 47, 977–996.

LITTLE, RODERICK J.A. (1979), "Discussion," in *Symposium on Incomplete Data: Preliminary Proceedings*, Washington, D.C.: U.S. Department of Health, Education, and Welfare, 290–295.

——— (1981), "Superpopulation Models for Nonresponse II: The Non-Ignorable Case," paper for Panel on Incomplete Data, Committee on National Statistics, National Academy of Sciences.

MINCER, JACOB (1974), *Schooling, Experience, and Earnings*, New York: National Bureau of Economic Research.

MORRIS, CARL N. (1979), "Nonresponse Issues in Public Policy Experiments, with Emphasis on the Health Insurance Study," in *Symposium on Incomplete Data: Preliminary Proceedings*, Washington, D.C.: U.S. Department of Health, Education, and Welfare, 448–470.

NELSON, FORREST D. (1977), "Censored Regression Models with Unobserved, Stochastic Censoring Thresholds," *Journal of Econometrics*, 6, 309–327.

OH, H. LOCK, and SCHEUREN, FREDERICK (1981), "Weighting Adjustments for Unit Nonresponse," paper for Panel on Incomplete Data, Committee on National Statistics, National Academy of Sciences.

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

ROYALL, RICHARD M., and HERSON, JAY (1973a), "Robust Estimation in Finite Populations I," *Journal of the American Statistical Association*, 68, 880–889.

——— (1973b), "Robust Estimation in Finite Populations II: Stratification on a Size Variable," *Journal of the American Statistical Association*, 68, 890–893.

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

——— (1978), "Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse," *Proceedings*, American Statistical Association Section on Survey Research Methods, 20–28.

——— (1979), "Discussion," in *Symposium on Incomplete Data: Preliminary Proceedings*, Washington, D.C.: U.S. Department of Health, Education, and Welfare, 285–289.

——— (1981), "Conceptual Issues in the Presence of Nonresponse," paper for Panel on Incomplete Data, Committee on National Statistics, National Academy of Sciences.

SCHAIBLE, WESLEY L. (1979), "Estimation of Finite Population Totals from Incomplete Sample Data: Prediction Approach," in *Symposium on Incomplete Data: Preliminary Proceedings*, Washington, D.C.: U.S. Department of Health, Education, and Welfare, 170–187.

TOBIN, JAMES (1958), "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26, 24–36.