MISSING DATA PROBLEMS IN COVERAGE EVALUATION STUDIES

Robert Fay and Charles Cowan, U.S. Bureau of the Census

1. Introduction

Since 1950 the U.S. Bureau of the Census has conducted major studies to evaluate the net or gross omissions of persons from the decennial censuses through direct survey methods. The 1980 Post-Enumeration Program (PEP) is the most recent member of this series. Typically, a sample survey of households has been used to provide the sample of persons for whom the existence of matching census enumerations is evaluated; in some instances, persons have been been selected from list frames based on administrative records instead.

In spite of considerable effort expended on these studies, they have historically been regarded as less than successful in achieving a measurement of the true error of the census. The principal evidence for this judgment rests with the systematically higher and more internally consistent estimates of net census error for previous censuses given by demographic analysis compared to the survey methods. Demographic analysis, which employs statistics on births, deaths, immigration, and emigration, as well as counts from earlier censuses and related sources, provided the Bureau's "preferred" methodology for the measure ment of the national undercount in 1970.

The limited success of direct survey methods in measuring the undercount has been attributed to the effect of "correlation bias", the tendency of persons actually missed from the census to be also disproportionately underrepresented in the sample used to evaluate the census. Particularly in the context of sample surveys, the same factors associated with omission of persons from the census may similarly lead to their omission from the effective coverage of the sample surveys.

Evaluations of census coverage by direct survey methods typically face a number of additional problems. (For a review see, for example, Cowan and Hogan, 1980.) One of these problems, the effect of missing or partial data, is the subject of this paper. The thesis presented here is that even when the overall level of nonresponse is low to moderate by normal standards of sample survey practice, there are often potentially complex links between the variable of interest, inclusion in the census, and the pattern of missing data. Although the issue of missing data does not supplant that of correlation bias as the preeminent limitation of survey methods to measure net census errors, it will be argued here that the issue of missing data is perhaps next most important in many of these studies.

A potential link between the pattern of nonresponse and an analytic variable of interest can arise in almost any survey context, but in survey evaluation of census coverage the linkage is almost assured by the procedural design of these studies. To illustrate this point, many studies can be summarized by the following simplified steps:

A sample of persons is selected and their probable census day address ascertained, then,

1. A majority of cases are (typically) matched to the census at the available addresses.

2. The remaining cases are sent to a "follow-up" procedure to use more intensive efforts, such as interviewing the persons themselves, to verify or correct the information on their census day address or to determine, in some instances, if they are out-of-scope (for example, not alive on census day). The actual outcomes might be divided into four groups:

a. Completed follow-up operation with the outcome of being matched to the census.

b. Completed follow-up operation with the outcome of not being matched to the census.

c. Completed follow-up operation with the outcome of being determined to be out-of-scope.

d. Incomplete follow-up operation.

In this simplified example, the last group, 2d, represents incomplete data, since these cases are not determined to be definitively matched or definitively non-matched to the census. (Of course, in actual cases additional sources of nonresponse may arise, such as obtaining no preliminary census day address for a subset of persons.)

Historically, one approach frequently used to address the question of missing data has been to analyze only the complete cases and to exclude the cases with incomplete data entirely. This procedure can be simply stated to rest on an assumption that the incomplete cases have the same non-matched rate as the complete cases. Generally speaking, the assumption that incomplete cases resemble complete cases is commonplace in survey sampling practice. The sometimes questionable nature of this assumption in this context, however, is more self-evident when examined from the perspective of persons in the survey population. In some applications, such as the 1980 PEP, factors operating largely independently of census coverage may contribute to the classification of these cases into category 2d, those with incomplete follow-up. For example, again in the case of the 1980 PEP, a large proportion of persons in the follow-up population (those persons not in category 1) that moved to a different address between the original sample survey and the follow-up survey almost a year later were classified into category 2d. The fact that these persons moved, particularly if the move was in the fall of 1980 or winter of 1981, would seem to be only tenuously linked, at most, to their enumeration in the census in April 1980 or shortly thereafter. Consider, then, persons in the original sample who moved after all census operations have

been completed and became unreachable to any potential follow-up operation. Most of these persons that had also been enumerated in the census would be readily matched to their census enumerations without the necessity for follow-up; thus, they would have been classified in category 1, matched before follow-up. The remaining persons enumerated in the census would then fall into category 2d. The persons missed from the census, on the other hand, would be all classified in category 2d. Thus, regardless of the proportion actually missed from the census, persons who are unreachable at the time of follow-up contribute only to the matched and incomplete cases. Except under the most extreme assumptions, the procedure of analyzing only the completed cases for this group is a self-evident source of bias in the estimated proportion of missed persons in the population.

The purpose of this paper is to establish a theoretical framework in which to discuss this problem and more complex problems of missing data in the studies of census coverage. The approach will be to suggest a synthesis of two more general areas of research. One of these is the growing methodological research into nonresponse in sample surveys. An important reference in this area, a paper by Little (1982), will be cited almost exclusively in the presentation here, because it summarizes or develops the applicable theory from this area of research that will be related in this paper to the general problems of nonresponse in studies of census coverage.

The second theoretical development to be cited here, the methodology of causal analysis for categorical data by Goodman (1972, 1973a, 1973b, 1978) is perhaps less obviously connected to the problem of missing data in coverage evaluation studies than the first. This second body of literature develops the correct applications of log-linear models to situations in which relationships among variables are structured by causal mechanisms. Census coverage and other related variables are generally categorical in nature, thus forming a basis for a possible application of this theory. More importantly, the issue of proper interpretations of the incomplete data from such studies rests on the mechanisms that are assumed to lead to the missing data. The causal models discussed by Goodman, particularly when supplemented with his later work in latent structure analysis, form an extremely rich class of models with potential application to a wide variety of assumptions and situations regarding the missing data.

The next section will suggest the implications of these ideas in some simple hypothetical situations, while the third section will discuss the relationship of these ideas to the imputation procedures designed for the 1980 PEP. The last section includes suggestions and observations about missing data in similar studies.

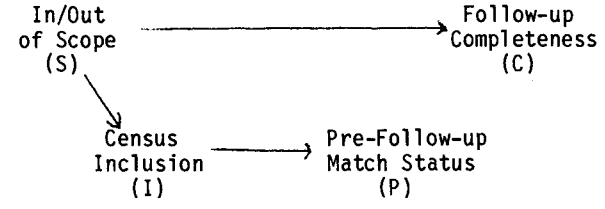
2. Two Simple Causal Models for Non-Response in Coverage Evaluation Surveys

In order to illustrate the implictions of the theory for missing data and for causal models cited in the preceding section, two models will be presented here that, in some contexts, may be applicable to the hypothetical general example outlined in the first section. The two models arise out of different specific assumptions about the mechanisms underlying the nonresponse.

Suppose that the original sample of persons includes a significant number who are in fact out of the scope of the analysis. Of course, some of these persons may be identified in the follow-up as belonging to category 2c defined in the first section, those determined to be out-of-scope. Suppose, however, that the original sample includes names with grossly incorrect addresses as a result of typographical or other clerical error, plus perhaps persons who had moved from the given address long before census day, and who, as a consequence of the passage of time, had become unknown to persons now at that address and to neighbors and other sources that might have been consulted. All such persons would be assumed not to match originally to the census, but to be divided between 2c, determined to be out-of-scope, and 2d, incomplete follow-up cases.

Suppose further that the follow-up effort is sufficiently intense that all in-scope persons may be classified into categories 1, 2a, or 2b. "Intense" is a necessary characterization here, for no other competing risks, such as movement of the persons to a different address after census day, must be allowed to interfere. In other words, if the persons themselves cannot be reached, rules on the acceptance of proxy information must be sufficiently broad to include all in-scope persons. The study essentially must presume that the consequent proxy information is of acceptable quality for the analysis.

Under these assumptions, group 2d represents a hidden group of out-of-scope cases, in spite of the fact that they may not have been explicitly determined to be so under the original follow-up rules. A causal model for these relationships is given by:



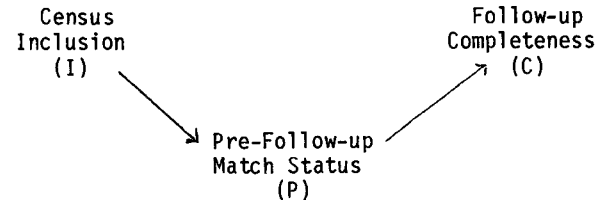
Under the stated assumptions, the cross-classification of these four variables contains a number of structural zeros. (These may be incorporated into the methodology for causal models by specifying the log-linear models in mulitiplicative rather than additive form.) Under these assumptions, the observed data actually satisfy

	In Scope		Out of Scope	
	In Census	Out of Census	In Census	Out of Census
Follow-up Complete				
Pre-Follow-up Match	1	0	0	0
Pre-Follow-up Non-Match	2a	2b	0	2c
Follow-up Incomplete				
Pre-Follow-up Match	0	0	0	0
Pre-Follow-up Non-Match	0	0	0	2d

In this circumstance, the response mechanism is decisively non-ignorable, as defined by Little (1982) and others since non-response implies logically that the case is out of scope, which in turn implies that the case is (by presumption) not enumerated in the census. In this case, the "completed cases only" analysis that ignores the responses in category 2d is appropriate, since this approach is equivalent to analyzing the cases in the scope of the evaluation.

This example illustrates a common situation, in which the ignorable response model studied in the theoretical literature, is not equivalent to the analysis of only complete cases occasionally reflected in some practice.

The second illustrative model makes diametrically opposite assumptions about the mechanisms underlying non-response. In this model, the importance of the out-of-scope persons is minimized, perhaps by an assumption that all such persons can be eliminated in the initial specification of the sample. Instead, the sample of person is viewed as comprising persons who are readily matched to census, persons who could be matched to the census if follow-up information were obtained, and persons not in the census. Further, it is assumed that mechanisms other than census coverage are responsible for incomplete follow-up information. More specifically, the causal model is



Census inclusion affects pre-follow-up match status because no excluded persons can be initially matched. Pre-follow-up match status affects follow-up completeness since all pre-follow-up matched persons (category 1 defined in the first section) are complete cases. The assumption of the causal model is that there is no additional effect of I on C that is not determined from the path through P.

The cross-classification consistent with this model is given by

	In Census	Out of Census
Follow-up Complete		
Pre-Follow-up Match	1	0
Pre-Follow-up Non-Match	2a	2b
Follow-up Incomplete		
Pre-Follow-up Match	0	0
Pre-Follow-up Non-Match	m_1	m_2

where m_1 and m_2 are missing data known to sum to 2d. Under this model, the Fisher-consistent estimator of m_1 is $2d \cdot 2a/(2a+2b)$. In other words, the model implies that the proportion of matchable cases among the incomplete cases is equal to that among the complete cases sent to follow-up. Generally, this model results in a higher proportion of estimated non-matches than the "completed cases only" analysis appropriate under the first model.

If, in the notation of Little (1982), variable u_s is taken to be the pre-follow-up match outcome, observed for all sample cases, and v_{sr} the final classification for the complete cases (that is, cases in group 1, 2a, or 2b), this second model is equivalent to an assumption of an ignorable response mechanism Here again, there is a distinction between ignorability of the response mechanism and the appropriateness of a "completed cases only" analysis of the data, since this is a case of an ignorable response mechanism that cannot be correctly treated by excluding the cases with incomplete data.

In cases where 2d is of any significant magnitude, even in relation to 2b, the analyses under the two different models will lead to substantially different results. This difference is not necessarily an indication that there is nothing to choose between them in given contexts. Since the models rest on markedly different assumptions, the natural question is whether one set of assumptions is more appropriate to the data at hand than another. Furthermore, each incorporates simplifying assumptions that could be modified by consideration of richer models. Section 4 will comment further on these questions.

3. Causal Model Underlying the Preliminary Treatment of Missing Data in the 1980 Post-Enumeration Survey

This section will describe a model which might be considered to form the essential rationale for the design of the imputation system for treatment of incomplete data in the estimation of gross omissions from the 1980 census. The discussion here will be in broad terms only, and several necessary refinements will not be presented in detail. It should be understood that these data are currently under review, and that the system described here represents only one of the methods to treat

the missing data that have been considered.

The sample for the PEP comprised the full April 1980 and August 1980 samples of the Current Population Survey (CPS), plus supplemental samples of persons in military barracks and of the institutional population. For purpose of analysis of missing data, each of the four samples was treated separately. Each of the samples was based on interviews of persons after census day, April 1, 1980, with April CPS interviews on the third week of April occurring closest to April 1. (An exception, not discussed here, occurs for an analysis of CPS households that were non-interviewers in April CPS.)

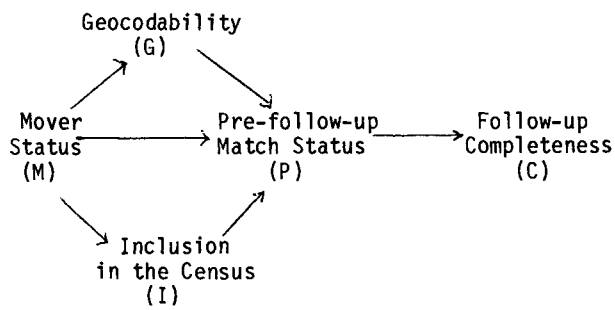
As in the hypothetical example in the first section of this paper, an initial match to the census enumeration was performed, with the result that over 85 percent of the sample was matched at this stage. Essentially, all remaining cases were designated for follow-up interview. (Exceptions to this rule constituted another special group not covered by the discussion here but which were treated by the imputation procedure. Essentially, such cases not sent to follow-up by design were imputed in the same manner as incomplete follow-up interviews.) The purpose of the follow-up interview was both to correct incomplete or erroneous information about the geographic location of the census day address, and in some cases to determine a different address to correct the address that had been searched originally. A significant proportion of cases sent to follow-up resulted in noninterviews, partially as a consequence of strict rules governing use of proxy information. (No use of information from neighbors was permitted, for example.)

The cases with complete follow-up interviews were classified into two groups: those that could be definitively classified into one enumeration district (ED) or an appropriately small group of ED's for searching, and those who could not. The latter group included very few cases of "nonmovers", persons with the same census day address as in the original sample, since the PEP interviewers were able to collect sufficient goegraphic information to establish the correct location of sampled housing units with respect to the geography of the 1980 census. The probelm came instead from persons who had moved between April 1, 1980 and the original interview; for such persons, the respondent's description was the only available resource to determine the ED in which each person should have been enumerated.

For persons who had been classified as complete interviews and for whom an area of search or single census enumeration had been determined, a series of clerical procedures were performed (although at the final stages by technical staff from the Census Bureau's Washington headquarters) to classify the case as either matched or missed. (An insignificant group of cases were classified as incomplete if materials were unavailable to complete the operations, and discussion of this problem will be omitted here.) Cases where the follow-up interview was incomplete and those for whom the 1980 census day address could not be assigned to an enumeration district or acceptable area of search formed two distinct groups of incomplete cases. The imputation procedure for PEP is closely linked to a causal model relating five categorical variables. Three of these appeared in the examples of the previous section: inclusion in the census (I), pre-follow-up match status (P), and follow-up completeness (C). Two other variables, mover status (M), and geocodability (G) are variables considered as a consequence of the timing and design of the PEP. Since CPS household interviews occurred after census day, part of the interviewed sample, termed the "movers," had a different address on April 1, 1980.

The concept of geocodability measures the ability to assign a case to the correct 1980 enumeration district (ED) or suitably small group of ED's prior to search. All cases matched before follow-up and cases with complete follow-up interviews that were assigned ED's for search are operationally defined to be geocodable; complete follow-up cases that could not be so assigned are not geocodable. For incomplete follow-up cases, the question of geocodability is not directly observed; rather it represents whether adequate geographic information would have been obtained had a follow-up interview been completed. Among completed cases, almost all that were not geocodable were also movers. In other words, geocoding could be completed for almost all cases at the location of the CPS interview, since CPS interviewers were able to provide detailed maps for the sampled housing units. The location of the census day address for movers, however, depended only upon information provided by respondents, and these cases naturally presented more challenging problems.

The following model summarizes a set of causal assumptions that might be made about the April CPS data:



In this model, mover status is allowed to affect inclusion in the census, because some empirical evidence suggests that persons who move near the time of the census are at higher risk of being omitted. Mover status is also allowed to affect geocodability, since for the most part, only movers are not geocodable. The model assumes that G and I are conditionally independent given M; in other words that there may be some marginal association between geocodability and inclusion in the census, but that this association is due to a common link to mover status.

Pre-follow-up matched status is a consequence of three variables, geocodability, mover status and inclusion in the census. In fact, all three are related in a logical way to M, since in the April

sample all ungeocodable cases, all census omissions, and all movers would appear among the pre-follow-up non-matches.

In the model, follow-up completeness is allowed to depend upon pre-follow-up match status (P), since all cases where P is "matched" are complete with respect to follow-up.

Associations between C and other variables are assumed to be through P. The application of this model to the April PEP sample implies a number of structural zeros and cells with missing data:

C=Complete	M=Non-movers		M=Movers	
G=geocodable	I=M	I=NM	I=M	I=NM
P=M	a	0	0	0
P=NM	b	c	d	e
G=non-geocodable				
P=M	0	0	0	0
P=NM	m_1	m_2	m_3	m_4
C=incomplete				
G=geocodable				
P=M	0	0	0	0
P=NM	m_5	m_6	m_7	m_8
G=non-geocodable				
P=M	0	0	0	0
P=NM	m_9	m_{10}	m_{11}	m_{12}

In this notation, the subscripted values of m are missing cells to be estimated. The marginals $m_1 + m_2$, $m_3 + m_4$, and $m_5 + m_6 + m_7 + m_8 + m_9 + m_{11} + m_{12}$ are available from the observed data.

This representation is a recursive causal model, again extending the methodology to reflect structural zeros by expressing the model in multiplicative rather than additive form. Furthermore, it is possible, although tedious, to compute closed-form expressions for the maximum likelihood estimates of all of the missing values under a multinomial sampling model. This is most easily accomplished in two steps:

a) The recursive causal model specifies an extended log-linear model for the five-way table reflecting the complete interaction M, G, I, and P, plus the effect of P on C. This model can be used to solve for m_5, m_6, m_7, m_8, $m_9 + m_{10}$, and $m_{11} + m_{12}$ in terms of b, c, d, e, $m_1 + m_2$, and $m_3 + m_4$. For example,

$$m_5 = b \cdot \frac{(m_5 + m_6 + m_7 + m_8 + m_9 + m_{10} + m_{11} + m_{12})}{(b + c + d + e + m_1 + m_2 + m_3 + m_4)} = b\,r$$

where the value of r is directly available from the data.

b) The recursive causal model implies an extended log-linear model for the four-way table of M, G, I, and P reflecting the interaction of M, G, and I and main effects of MP, GP and IP. The pattern of structural zeros conforms to this model. The recursive causal model also implies a log-linear model for the three-way table of M, G, and I including MG and MI main effects. This model can be used to solve for m_1, m_2, m_3, m_4, m_9, m_{10}, m_{11}, and m_{12}, taking advantage of the given marginals $m_1 + m_2$ and $m_3 + m_4$ and the estimates of $m_9 + m_{10}$ and $m_{11} + m_{12}$ from the preceding step. For the resulting three-way table, for example,

$$(1+r)\,m_1 = \frac{a + b(1+r)}{a + b(1+r) + c(1+r)} \cdot (1 + r)\,(m_1 + m_2)$$

Besides being maximum-likelihood, these estimates are the unique Fisher-consistent solution to this problem, so they are also appropriate even when the sampling distribution is not multinomial.

This model is an interesting instance of a non-ignorable response mechanism. The variables form a monotone response situation, as described by Little (1982), since P and C are measured for all, M and G are measured for C=complete, and I is measured only for C=complete and G=geocodable. In the two-step calculation of the maximum likelihood estimates under the causal model, step a is consistent with an assumption of an ignorable response mechanism for non-interview on follow-up. The second step is not consistent with such a model, however; the collapsing to the three-way table across variables P and C, required by the recursive causal model, is inconsistent with an ignorable response model.

The preliminary imputation system designed for PEP paralleled the logic of the preceding causal model. In place of the five-way table just discussed, a complex statistical matching procedure was implemented in order to attempt to represent the effect of a number of additional covariates. The statistical match was performed in two waves:

a) Cases with incomplete follow-up interviews were imputed to mover status, M, and geocoding status, G. If the statistically matched case had a value of G of geocoded, the resulting value of I was also imputed. The incomplete cases in this wave were statistically matched to cases with complete follow-up interviews who were sent to follow-up, i.e, who were not initially matched to the census before follow-up. This exclusion corresponds to the omission of cell "a" from the first step of calculation of the maximum likelihood estimates.

b) A second wave of imputation was used to im-

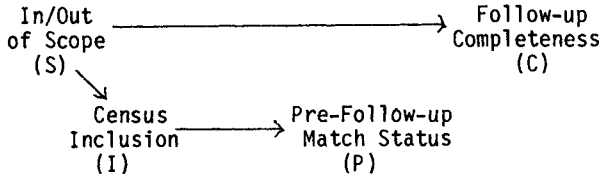
pute the value of I for those classified as not geocodable, including those imputed to this classification in the previous step. In this instance, the pre-follow-up codes were ignored in the statistical match, thus allowing cell "a" into the computation, as would be required by the causal model. (At this stage, weighting adjustments were substituted for actual imputation for some classes of cases with the intent of producing an equivalent effect.)

Thus, causal analysis in this instance provides a model for a non-ignorable response mechanism that nonetheless can be estimated from the data, and a rationale for what would otherwise be an obscure imputation procedure. Furthermore, a fruitful approach to examine the potential limitation of the imputation procedure would be in terms of alternative causal models. For example, the effect of different associations between variables excluded from the preliminary model could be postulated and their implications studied. Further work on this question is planned.

4. Use of Other Causal Models for Analysis of Incomplete Data from Coverage Studies

The example in the previous section illustrated the application of a causal model in a situation where multiple factors were directly and indirectly related to response. The procedural design of the PEP permitted estimation of the model, so that the parameters were identifiable. Although PEP contained some distinct features that differ from many other studies of the sort, the possible effect of multiple sources of nonresponse is shared by most such studies. An appropriate objective of design, therefore, is the ability to estimate the effect and implications of different sources of nonresponse.

To illustrate this point, the hypothetical design of section 2 will again be examined. In the original design, two models were proposed, based on different presumed mechanisms underlying the data. By an appropriate alteration of the design, however, the separate effects of both sources of nonresponse can be estimated under an integrated model. The required change is to send the originally matched cases (actually, a managably small sample of them) to to follow-up in order to estimate the follow-up response function. The causal model for this integrated approach would be formally the same as that of model 1



In model 2, an effect of P on C is necessary as a consequence sending to follow-up only the preliminary non-matches. In the revised design, an association is potentially absent. The data would be interpreted as

	In Scope		Out of Scope	
	In Census	Out of Census	In Census	Out of Census
Follow-up Complete				
Pre-Follow-up Match	1a	0	0	0
Pre-Follow-up Non-Match	2a	2b	0	0
Follow-up Incomplete				
Pre-Follow-up Match	1d	0	0	0
Pre-Follow-up Non-Match	m_1	m_2	0	m_3

Only the total $m_1 + m_2 + m_3$ would be directly observed, but each of the three three components is estimable under the model. For example,

$$m_1 = 2a\ 1d/1a$$

The estimate of m_3 is derived as a final step by subtraction.

References

Goodman, Leo A. (1972), "A General Model for the Analysis of Surveys," American Journal of Sociology, 77, 1035-86.

(1973a), "Causal Analysis of Data from Panel Studies and Other Kinds of Surveys" American Journal of Sociology, 78, 1135-91.

(1973b), "The Analysis of Multidimensional Contingency Tables When Some Variables Are Posterior to Others: A Modified Path Analysis Approach," Biometrika, 59, 579-96.

(1978), Analysing Qualitative/Categorical Data, Cambridge, MA: ABT Associates Inc.

Hogan, Howard, and Cowan, Charles (1980), "Imputations, Response Errors, and Matching in Dual System Estimation," Procedings of the Social Statistics Section, 1980, American Statistical Association.

Little, Roderick J.A. (1982), "Models for Nonresponse in Sample Surveys," Journal of the Amer- Statistical Association", 77, 237-250.