CENSORING OF OUTCOMES AND REGRESSORS DUE TO SURVEY NONRESPONSE: IDENTIFICATION AND ESTIMATION USING WEIGHTS AND IMPUTATIONS

Joel L. Horowitz
Department of Economics
University of Iowa
Iowa City, Iowa 52242

and

Charles F. Manski
Department of Economics
University of Wisconsin-Madison
Madison, Wisconsin 53706

October 1995

We thank Tue Gorgens for research assistance. The research of Joel Horowitz was supported in part by National Science Foundation Grants DMS-9208820 and SBR-9307677. The research of Charles Manski was supported in part by National Science Foundation Grant SBR-9223220. An early draft of some of the material in this paper was circulated during 1994 with the title "Joint Censoring of Regressors and Outcomes: Survey Nonresponse and Attrition."

Abstract

Survey nonresponse makes identification of population statistics problematic. Except in special cases, identification is possible only if one makes untestable assumptions about the distribution of the missing data. However, non-response does not preclude identification of bounds on population statistics. This paper shows how identified bounds on unidentified population statistics can be obtained under several forms of nonresponse. Organizations conducting major surveys commonly release public-use data files that provide nonresponse weights or imputations to be used for estimating population statistics. The paper shows how to bound the asymptotic bias of estimates using weights and imputations. The results are illustrated with empirical examples based on the National Longitudinal Survey of Youth.

1. Introduction

Survey nonresponse is problematic for the identification of population statistics. Whether nonresponse takes the form of particular missing data items or entire missing interviews, the only way to identify population statistics is to make assumptions that determine the distribution of the missing data. A fundamental problem of empirical analysis is that such assumptions are untestable.

Organizations conducting major surveys commonly release public-use data files that provide nonresponse weights and imputations to be used for estimating population statistics. Weights are also used to compensate for planned variation in sampling rates across strata of the population, but this usage is distinct and not of concern here. Available survey data cannot validate the distributional assumptions underlying weights and imputations. These data can, however, be used to determine the logically possible values of the population statistic of interest and, hence, to bound the asymptotic bias of estimates using weights and imputations. This paper shows how.

Our analysis is unabashedly conservative. Our objective is to characterize the inferences that can be drawn in the absence of untestable assumptions about the distribution of the missing data. Through this "worst-case" analysis, we hope to correct the imbalance that we perceive in the literature on survey nonresponse. The literature has focused almost entirely on "best-case" scenarios in which population statistics are identified by imposing strong assumptions on

the distribution of the missing data. For example, the *multiple imputation* analysis of Rubin (1987) assumes that nonresponse is ignorable. Likewise, the weights commonly provided with public-use data files are computed under the assumption that nonresponse is ignorable. As we see matters, it is not enough for empirical researchers to know that imputations and weights work well if specified assumptions hold. It is equally important to be able to characterize the inferences that may be drawn without imposing these assumptions. As emphasized in Manski (1995), conservative analysis of the type performed here allows the establishment of a domain of consensus among researchers who may hold disparate beliefs about what assumptions are appropriate.

Throughout the paper, we assume that each member of the population is characterized by a value of (y, x, w, z), where $y \in Y$ is an outcome of interest, $x \in X$ and $w \in W$ are covariate vectors, and z is a binary variable indicating nonresponse. We focus on the problem of estimating conditional expectations of the form $E[g(y) | x \in A]$, where $g(\cdot)$ is a specified real-valued function of the outcome y and A is a specified set of values of the covariates x. We assume that a random sample is drawn and that some observations of y and/or x are censored by nonresponse. We let $z = 1$ indicate that (y, x) is fully observed and $z = 0$ that data are missing. The covariate vector w, which is always observed, is used to compute weights and imputations.

In the theoretical sections of the paper, we focus attention on the core problem of identification and abstract from concerns of finite-sample inference.

Thus, we assume that features of the population identified by the censored random sampling process are known rather than simply estimable. When we speak of bias, we mean asymptotic bias. In several empirical examples, we examine both the identification and the sampling aspects of inference.

To introduce themes, Section 2 considers the familiar case in which the outcomes y are censored but the regressors x are not. In this setting, the identifiability of conditional expectations is transparent. We determine the probability limits of estimates using weights and imputations and bound the asymptotic bias of these estimates. We give an empirical example using data from the 1979 National Longitudinal Survey of Youth (NLSY79) to estimate employment probabilities.

Section 3 examines joint censoring of the outcomes y and the regressors x. We find that this problem, which has not been studied previously, is formally equivalent to the problem of identification when only outcomes are censored, except that the *effective* censoring rate is higher than the actual one. We give an empirical example using NLSY79 data to estimate unemployment rates.

Section 4 examines the case in which regressors are censored but outcomes are not. This problem, which also has not been studied previously, turns out to be related to the problem of identification from contaminated data studied in Horowitz and Manski (1995).

The three problems examined in Sections 2 through 4 -- outcome censoring, joint censoring, and regressor censoring -- are polar cases. In practice,

researchers often face mixed cases of greater complexity. Section 5 examines a type of mixed censoring problem that arises commonly in panel data. Some sample members are never interviewed, and so have their outcomes and all regressors censored. Others are interviewed in early waves of the survey but later drop out of the panel, and so have their outcomes and a subset of regressors censored. We use the NLSY79 data to illustrate.

The analysis in Sections 2 through 5 takes the administration of a survey as given. Section 6 considers the situation of a survey organization which may act to reduce the nonresponse rate. Here we characterize how increases in response rates yield gains in identification. Section 7 gives conclusions.

An important generic finding of this paper is that estimates using weights are potentially more biased than are those using imputations. Estimates using imputations take the observed data as given and specify logically possible values for the missing data. Hence imputation always yields a logically possible value of the conditional expectation of interest, regardless of how the imputed values are computed. Estimates using weights modify the observed data. We show that, unless the weights are computed in particular ways, weighting need not yield a logically possible value of the conditional expectation.

2. Outcome Censoring

We begin with the widely appreciated and easily understood problem of outcome censoring, also known as the *selection problem*. Thus, in this section we assume that only observations of y are missing. Realizations of x are always observed. The binary variable z indicates whether or not y is observed.

2.1. Identification and the Asymptotic Bias of Estimates

We begin with the identifiability of $E[g(y)|x \in A]$. As in Manski (1994, 1995), let $K_0 \equiv \inf_{y \in Y} g(y)$ and $K_1 \equiv \sup_{y \in Y} g(y)$, where Y is the domain of y. Define $E[g(y)|A] \equiv E[g(y)|x \in A]$ and observe that

$$(1) \quad E[g(y)|A] = E[g(y)|A, z = 1]\cdot P(z = 1|A) + E[g(y)|A, z = 0]\cdot P(z = 0|A).$$

The censored sampling process identifies the mean $E[g(y)|A, z = 1]$ of the uncensored observations, the response probability $P(z = 1|A)$, and the nonresponse probability $P(z = 0|A)$, but provides no information on the censored mean $E[g(y)|A, z = 0]$. The last quantity, however, necessarily lies in the interval $[K_0, K_1]$. This implies the sharp bound

$$(2) \quad E[g(y)|A, z = 1]\cdot P(z = 1|A) + K_0\cdot P(z = 0|A) \leq E[g(y)|A] \leq E[g(y)|A, z = 1]\cdot P(z = 1|A) + K_1\cdot P(z = 0|A).$$

Thus, the available survey data do not identify $E[g(y)|A]$ but do restrict its value to an interval of width $(K_1 - K_0)\cdot P(z = 0|A)$.

Let θ_N denote an estimate of $E[g(y)|A]$ obtained in a sample of size N and let θ be its probability limit. The bias of θ (or asymptotic bias of θ_N) is $\theta - E[g(y)|A]$. By (2),

$$(3)\quad \theta - E[g(y)|A, z = 1]\cdot P(z = 1|A) - K_1\cdot P(z = 0|A) \le \theta - E[g(y)|A] \le \theta - E[g(y)|A, z = 1]\cdot P(z = 1|A) - K_0\cdot P(z = 0|A).$$

The lower bound in (3) is the maximum negative bias of θ and the upper bound is the maximum positive bias.

An empirical researcher should restrict attention to estimates θ that lie within the bound (2) on $E[g(y)|A]$. After all, if θ lies outside the bound, an estimate with unambiguously smaller bias is always available, namely the bound endpoint nearer to θ.

It is relevant to note that over forty years ago, in a study of the statistical problems of the Kinsey report on sexual behavior, Cochran, Mosteller, and Tukey (1954, pages 274 - 282) used bounds of the form (2) to express the possible effects of nonresponse to the Kinsey survey. Unfortunately, the subsequent literature did not pursue the idea.

2.2. Estimation Using Weights and Imputations

WEIGHTS: Suppose that a random sample of size N has been drawn. Let N(1, A) denote the subset of sample members for whom z = 1 and x ϵ A. In weighting approaches to estimation of E[g(y)|A], one assigns to each member of N(1, A) a weight s(w) that varies with the observed value of the covariate w. One then estimates E[g(y)|A] by the weighted average

$$(4) \quad \theta_N \equiv \frac{1}{N_{1A}} \sum_{i \in N(1, A)} s(w_i) \cdot g(y_i),$$

where N_{1A} is the number of sample members in N(1, A). The probability limit of θ_N is

$$(5) \quad \theta \equiv E[s(w) \cdot g(y) | A, z = 1].$$

A minimal criterion for reasonableness of any estimate of E[g(y)|A] is that the estimate should asympotically lie within the bound (2), which gives the logically possible values of E[g(y)|A]. Weighted means of the form (5) necessarily satisfy this criterion if the weights s(·) have particular forms, but may not satisfy it otherwise.

The public-release files of major surveys typically provide weights of the form

$$(6)\quad s(w) \;=\; \frac{P(w)}{P(w \mid z = 1)} \;=\; \frac{P(z = 1)}{P(z = 1 \mid w)}\,.$$

A simple example suffices to show that estimates using such weights can asymptotically lie outside the logically possible range of $E[g(y) \mid A]$. Suppose that nonresponse on y is concentrated outside the set A; that is, $P(z = 1) < 1$ but $P(z = 1 \mid A) = 1$. Then the bound (2) is degenerate at the value $E[g(y) \mid A, z = 1]$. The weighted estimate lies outside the bound unless the estimate happens to equal $E[g(y) \mid A, z = 1]$.

The basic flaw in using public-use weights of the form (6) to estimate conditional means of the form $E[g(y) \mid A]$ is that the response probabilities $P(z = 1)$ and $P(z = 1 \mid w)$ do not condition on the event $[x \in A]$. This flaw is eliminated by use of weights of the form

$$(7)\quad s(w) \;=\; \frac{P(z = 1 \mid A)}{P(z = 1 \mid A, w)}\,.$$

With these weights, Bayes Theorem implies that

$$(8)\quad s(w) \cdot P(w \mid A, z = 1) \;=\; P(w \mid A).$$

It follows that

$$(9)\quad E[s(w)\cdot g(y)\,|\,A,\ z=1]$$

$$= \sum_{v\in W} E[s(w)\cdot g(y)\,|\,A,\ z=1,\ w=v]\cdot P(w=v\,|\,A,\ z=1)$$

$$= \sum_{v\in W} E[g(y)\,|\,A,\ z=1,\ w=v]\cdot P(w=v\,|\,A)$$

$$= \sum_{v\in W} E[g(y)\,|\,A,\ z=1,\ w=v]\cdot P(w=v\,|\,A,\ z=1)\cdot P(z=1\,|\,A)$$

$$+ \sum_{v\in W} E[g(y)\,|\,A,\ z=1,\ w=v]\cdot P(w=v\,|\,A,\ z=0)\cdot P(z=0\,|\,A)$$

$$= E[g(y)\,|\,A,\ z=1]\cdot P(z=1\,|\,A)$$

$$+ \sum_{v\in W} E[g(y)\,|\,A,\ w=v,\ z=1]\cdot P(w=v\,|\,A,\ z=0)\cdot P(z=0\,|\,A).$$

Observe that $K_0 \le E[g(y)\,|\,A,\ w=v,\ z=1] \le K_1$, all $v \in W$. Hence the right side of (9) lies within the bound (2).

Selection of weights of the form (7) implies that the asymptotic weighted estimate $E[s(w)\cdot g(y)\,|\,A,\ z=1]$ is a logically possible value of $E[g(y)\,|\,A]$, but does not imply that this estimate equals $E[g(y)\,|\,A]$. Inspection of the right side of (9) shows that the estimate does equal $E[g(y)\,|\,A]$ if $g(y)$ is mean-independent of z, conditional on (A, w); that is, if

$$(10)\quad E[g(y)\,|\,A,\ w,\ z=1] = E[g(y)\,|\,A,\ w,\ z=0].$$

If this untestable condition holds, nonresponse on g(y) is said to be *exogenous* or *ignorable* conditional on (A, w).

An interesting result holds whenever the covariates w used to form weights of the form (7) are statistically independent of y, conditional on the event (A, z = 1). In such cases, (5), (7), and (8) imply that

$$
\begin{aligned}
(11)\quad \theta &= E[g(y)\,|\,A,\ z = 1]\cdot E[s(w)\,|\,A,\ z = 1] \\
&= E[g(y)\,|\,A,\ z = 1] \sum_{v \in W} s(v)\cdot P(w = v\,|\,A,\ z = 1) \\
&= E[g(y)\,|\,A,\ z = 1] \sum_{v \in W} P(w = v\,|\,A) \\
&= E[g(y)\,|\,A,\ z = 1].
\end{aligned}
$$

Thus the weighted estimate reduces to the unweighted one $E[g(y)\,|\,A,\ z = 1]$. This result may explain some of the many reports by empirical researchers that their weighted and unweighted estimates of population statistics are similar in magnitude.

IMPUTATION: In imputation approaches, one assigns equal weights to all respondents, whether or not they report y. Each sample member who does not report y is assigned an imputed value $y^* \in Y$. The quantity $E[g(y)\,|\,A]$ is estimated by the sample average

$$(12)\quad \theta_N = \frac{1}{N_A} \sum_{i \in N(A)} [g(y_i)z_i + g(y_i^*)(1 - z_i)],$$

where N(A) is the subsample of respondents with $x \in A$, and N_A is the size of N(A). The probability limit of this estimate is

$$(13)\quad \theta \equiv E[g(y)|A, z = 1]\cdot P(z = 1|A) + E[g(y^*)|A, z = 0]\cdot P(z = 0|A).$$

Imputation uses the observed data as given and directly imposes a distribution on the censored data. Hence estimates using imputation always yield logically possible values of $E[g(y)|A]$, regardless of how the imputed values are computed. To verify this, observe that $K_0 \leq E[g(y^*)|A, z = 0] \leq K_1$. Hence the right side of (13) necessarily lies within the bound (2).

Comparison of (13) and (1) shows that an estimate using imputation equals $E[g(y)|A]$ if

$$(14)\quad E[g(y)|A, z = 0] = E[g(y^*)|A, z = 0].$$

The available data do not reveal the left side of (14), so this condition is untestable.

Imputation requires a rule to generate the predicted values y^*. It is common to assign each respondent i with missing outcome data the imputed value

(15) $y_i^* = E(y|A, w = w_i, z = 1)$.

The resulting estimate equals $E[g(y)|A]$ if $g(\cdot)$ is linear in y and nonresponse is exogenous. In multiple imputation, realizations of y_i^* are drawn at random from the distribution $P(y|A, w = w_i, z = 1)$. The probability limit of an estimate obtained in this way equals $E[g(y)|A]$ if y is statistically independent of z, conditional on (A, w).

2.3. An Empirical Example

This section presents an empirical example that illustrates some of the theoretical results derived in sections 2.1 and 2.2. We use the 1979 National Longitudinal Survey of Youth to estimate the probability that a member of the surveyed population is employed in 1991. The surveyed population consists of individuals born between January 1, 1957 and December 31, 1964 who resided in the United States in 1979. From 1979 on, the NLSY79 has yearly sought to obtain interviews with a random sample of this population and with several supplemental samples of special subpopulations (see Center for Human Resource Research, 1992). We use the random sample data here.

In this example, the outcome y indicates an individual's employment status at the time of the 1991 interview. In the 1979 base year, the NLSY79 sought to interview a random sample of 6812 individuals and succeeded in obtaining

interviews from 6111 of the sample members. Data on employment status in 1991 are available for 5556 of the 6111 individuals interviewed in the base year. The remaining 555 are nonrespondents, some because they declined to be interviewed in 1991 and some because they did not answer the employment-status question in their 1991 interviews. Table 1 presents these response statistics and the frequencies with which different outcome values are reported.

The population nonresponse rate, which takes account of sample members who were never interviewed, is $P(z = 0) = 1256/6812 = 0.184$. Researchers computing nonresponse rates to questions in the later years of longitudinal surveys often condition on the event that a sample member was interviewed in the base year. Let this event, which is always observed, be denoted BASE. Then the "ever-interviewed" nonresponse rate for employment status in 1991 is $P(z = 0 \mid \text{BASE}) = 555/6111 = 0.091$.

Table 1: 1991 EMPLOYMENT STATUS OF NLSY79 RESPONDENTS

Employment Status	Number of Respondents
Employed (y = 2)	4332
Unemployed (y = 1)	297
Out of Labor Force (y = 0)	927
Ever-interviewed Nonrespondents	555
Never-interviewed Nonrespondents	701
Total	6812

The empirical probability of employment among the 5556 individuals who respond to the 1991 employment-status question is $P(y = 2 \mid z = 1) = 4332/5556 = 0.780$. The censored probability of employment among nonrespondents is bounded by 0 and 1. Hence equation (2) yields the following bounds on the population and the ever-interviewed employment probabilities $P(y = 2)$ and $P(y = 2 \mid \text{BASE})$:

$$(0.780)(0.816) \leq P(y = 2) \leq (0.780)(0.816) + (0.184)$$
$$(0.780)(0.909) \leq P(y = 2 \mid \text{BASE}) \leq (0.780)(0.909) + (0.091)$$

or $0.636 \leq P(y = 2) \leq 0.821$ and $0.709 \leq P(y = 2 \mid \text{BASE}) \leq 0.800$.

SAMPLING VARIATION: The theoretical focus of this paper is identification, but empirical research must also be concerned with sampling variation. A useful way to characterize the implications of sampling variation is to present confidence intervals on the estimates of the bounds. Consider $P(y = 2)$. Some algebra shows that the bound given in (2) may be rewritten as

$$P(y = 2, z = 1) \leq P(y = 2) \leq 1 - P(y \neq 2, z = 1).$$

The asymptotic standard errors of the estimates of the lower and upper bounds are

$$S_L = \{P(y = 2, z = 1)[1 - P(y = 2, z = 1)]/N\}^{1/2}$$

and

$$S_U = \{P(y \neq 2, z = 1)[1 - P(y \neq 2, z = 1)]/N\}^{1/2}$$

where N = 6812 is the sample size. A Bonferroni asymptotic joint confidence region with level at least 95 percent is obtained by forming the intersection of individual 97.5 percent regions. These regions are the point estimates of the lower and upper bounds $\pm (2.24)S_L$ and $\pm (2.24)S_U$ respectively.

Substituting sample frequencies for population probabilities yields

$$P(y = 2, z = 1) = 4332/6812 = 0.636$$

$$P(y \neq 2, z = 1) = 1224/6812 = 0.180$$

$$S_L = [(0.636)(1 - 0.636)(1/6812)]^{1/2} = 0.0058$$

$$S_U = [(0.180)(1 - 0.180)(1/6812)]^{1/2} = 0.0047.$$

So the estimated asymptotic joint Bonferroni 95 percent intervals are

$$0.623 \leq \text{lower bound on } P(y = 2) \leq 0.649$$

$$0.810 \leq \text{upper bound on } P(y = 2) \leq 0.831.$$

Analogous computations conditioning on the event BASE yields

$$0.696 \leq \text{lower bound on } P(y = 2 \mid \text{BASE}) \leq 0.722$$

$$0.788 \leq \text{upper bound on } P(y = 2 \mid \text{BASE}) \leq 0.811.$$

Observe that these confidence intervals are much narrower than the widths of the estimated bounds. Thus, identification is the dominant problem in inference on P(y = 2) and P(y = 2|BASE) from the NLSY79 data; sampling variation is a second-order concern. Abstracting from the example, this is the generic

conclusion to be drawn except in situations where the sample size is quite small or the response rate is very close to one. Indeed, this conclusion was reached over forty years ago by Cochran, Mosteller, and Tukey (1954) in their examination of the implications of nonresponse in the Kinsey Survey.

WEIGHTED ESTIMATION: Finally, let us use the sampling weights provided with the NLSY79 public-use file to estimate the population employment probability. As described by Frankel, McWilliams, and Spencer (1983), these weights are intended for use in estimating population statistics from the "full" NLSY79 data, which consist of the random sample used here plus the supplemental samples of special subpopulations. Nevertheless, researchers often apply the weights to the random sample NLSY79 data. Insertion of the NLSY79 sampling weights into equation (4) yields 0.796 as the weighted estimate of the population employment probability.

The weighted estimate lies within the bounds and so gives a logically possible value for P(y = 2). The estimate implies that nonrespondents have higher employment probabilities than respondents. To see this, consider the probability decomposition

$$P(y = 2) = P(y = 2 \mid z = 1)P(z = 1) + P(y = 2 \mid z = 0)P(z = 0).$$

The raw NLSY79 data give P(z = 1) = 0.816 and P(y = 2 | z = 1) = 0.780. Using the weighted estimate 0.796 as the value of P(y = 2), the implied employment

probability for nonrespondents is $P(y = 2 \mid z = 0) = 0.867$. Thus, a researcher who applies the NLSY79 sampling weights implicitly assumes that the employment probability for nonrespondents is 0.087 higher than that for respondents. This result contrasts with the finding of MaCurdy, Gritz, and Mroz (1994) that individuals who subsequently dropped out of the panel and individuals who dropped out and later returned respectively had lower pre-drop out and post-return employment rates than individuals who never dropped out.

3. Joint Censoring of Outcomes and Regressors

Whereas we assumed in Section 2 that only y is censored, we examine here the joint censoring problem in which nonresponse on y is accompanied by nonresponse on x. Thus $z = 1$ now indicates that (y, x) is observed and $z = 0$ that (y, x) is missing.

Joint censoring of outcomes and regressors is the generic consequence of interview nonresponse, which occurs when sample members refuse to participate in the survey or cannot be contacted by survey administrators. Interview nonresponse is ubiquitous in survey research. Hence, it is curious that the identification problem generated by joint censoring has not previously been explored.

Joint censoring is also induced by conditioning on sets of outcome values.

Suppose that one wants to estimate a conditional expectation of the form $E[g(y)|y \in B]$, where B is a specified set of outcome values. When data are missing on the outcome y, they are also missing on the conditioning event $\{y \in B\}$. To illustrate, let y denote income, let y_0 denote the official poverty line, and consider the problem of inference on the mean income of persons below the poverty line. If some respondents do not report their incomes, inference on $E(y|\ y < y_0)$ poses a joint censoring problem.

3.1. Identification and the Asymptotic Bias of Estimates

To determine the identifiability of $E[g(y)|A]$, we begin again with the decomposition given in (1), namely

$$E[g(y)|A] = E[g(y)|A, z = 1]\cdot P(z = 1|A) + E[g(y)|A, z = 0]\cdot P(z = 0|A).$$

When only y is censored, the only unknown quantity on the right side is $E[g(y)|A, z = 0]$. When (y, x) are jointly censored, the response probability $P(z = 1|A)$ is not known either. We can, however, establish a sharp bound on $P(z = 1|A)$.

Use Bayes Theorem to write

$$(16)\quad P(z = 1|A) = \frac{P(A|z = 1)P(z = 1)}{P(A|z = 1)P(z = 1) + P(A|z = 0)P(z = 0)}.$$

Inspect the right side of (16). The sampling process identifies $P(A|z = 1)$, $P(z = 1)$, and $P(z = 0)$, but is uninformative about $P(A|z = 0)$. Setting $P(A|z = 0) = 1$ yields the sharp lower bound on $P(z = 1|A)$, and setting $P(A|z = 0) = 0$ yields the sharp upper bound. The result is

$$(17)\quad P_e(z = 1|A) \leq P(z = 1|A) \leq 1,$$

where

$$(18)\quad P_e(z = 1|A) \equiv \frac{P(A|z = 1)P(z = 1)}{P(A|z = 1)P(z = 1) + P(z = 0)}.$$

We call $P_e(z = 1|A)$ the *effective response probability* and $P_e(z = 0|A) = 1 - P_e(z = 1|A)$ the *effective nonresponse probability*.

We can now establish the identifiability of $E[g(y)|A]$. The censored mean $E[g(y)|A, z = 0]$ lies in the interval $[K_0, K_1]$ and the response probability $P(z = 1|A)$ lies in the interval $[P_e(z = 1|A), 1]$. Hence, (1) implies the sharp bound

$$(19)\quad E[g(y)\,|\,A,\ z=1]\cdot P_e(z=1\,|\,A) + K_0\cdot P_e(z=0\,|\,A) \ \le\ E[g(y)\,|\,A] \ \le\ E[g(y)\,|\,A,\ z=1]\cdot P_e(z=1\,|\,A) + K_1\cdot P_e(z=0\,|\,A).$$

The bound (19) under joint censoring of (y, x) has the same form as the bound (2) under censoring of y alone, except that the effective response probability replaces the actual response probability. Observe how the magnitude of the effective response probability varies with $P(A\,|\,z=1)$, falling from $P(z=1)$ to zero as $P(A\,|\,z=1)$ falls from one to zero. Thus, inference on $E[g(y)\,|\,A]$ becomes increasingly difficult as $P(A\,|\,z=1)$ falls, and is not feasible at all when $P(A\,|\,z=1)=0$. In this sense, joint censoring generates a qualitatively more difficult inferential problem than does outcome censoring.

Finally, (19) implies this sharp bound on the asymptotic bias of an estimate of $E[g(y)\,|\,A]$:

$$(20)\quad \theta - E[g(y)\,|\,A,\ z=1]\cdot P_e(z=1\,|\,A) - K_1\cdot P_e(z=0\,|\,A) \ \le\ \theta - E[g(y)\,|\,A] \ \le\ \theta - E[g(y)\,|\,A,\ z=1]\cdot P_e(z=1\,|\,A) - K_0\cdot P_e(z=0\,|\,A).$$

3.2. Estimation Using Weights and Imputations

WEIGHTS: When outcomes and regressors are jointly censored, the weighted average estimate (4) remains computable and its probability limit remains

$E[s(w)\cdot g(y)|A, z = 1]$. Nevertheless, there are important differences between the present situation and the one examined in Section 2.

The maximum asymptotic bias of the estimate is greater under joint censoring than under outcome censoring. Under outcome censoring, the available sample information implied that $E[g(y)|A]$ lies within the bound (2), so that bound was used to compute the bound (3) on the asymptotic bias of the estimate. Under joint censoring, the available sample information implies only that $E[g(y)|A]$ lies within the wider bound (19). Hence, we obtain the wider bound (20) on the bias of the estimate.

In Section 2, we saw that the weighted estimate asymptotically must lie within the bound (2) if the weights have the form (7), namely

$$s(w) = \frac{P(z = 1|A)}{P(z = 1|A, w)} .$$

Under joint censoring, the ratio $P(z = 1|A)/P(z = 1|A, w)$ is generally not identified and so weights of this form are not computable. An exception is the special case where w is statistically independent of z, conditional on A. Then (7) reduces to $s(w) = 1$ and we obtain the unweighted estimate $E[g(y)|A, z = 1]$.

IMPUTATION: When outcomes and regressors are jointly censored, the imputation estimate (12) considered in Section 2 is not computable because the composition of the subsample N(A) is not known. In the present setting, each sample member who does not report (y, x) may be assigned an imputed pair of values (y^*, x^*). The quantity $E[g(y)|A]$ may then be estimated by the sample average

$$(21)\quad \theta_N = \frac{1}{N_{1A} + N^*_{0A}} \sum_{i \in N(1, A)} g(y_i) + \sum_{i \in N^*(0, A)} g(y^*_i)$$

$$= \pi_N \frac{1}{N_{1A}} \sum_{i \in N(1, A)} g(y_i) + (1 - \pi_N) \frac{1}{N^*_{0A}} \sum_{i \in N^*(0, A)} g(y^*_i).$$

Here N(1, A) is the subsample of respondents with $z = 1$ and $x \in A$, $N^*(0, A)$ is the subsample with $z = 0$ and $x^* \in A$, and $\pi_N \equiv N_{1A}/(N_{1A} + N^*_{0A})$.

The probability limit of this estimate is

$$(22)\quad \theta \equiv E[g(y)|A, z = 1]\cdot\pi + E[g(y^*)|x^* \in A, z = 0]\cdot(1 - \pi),$$

where

$$(23)\quad \pi = \frac{P(A|z = 1)P(z = 1)}{P(A|z = 1)P(z = 1) + P(x^* \in A|z = 0)P(z = 0)}.$$

The value of π depends on the imputation rule but necessarily lies in the interval $[P_e(z = 1|A), 1]$. It follows that θ always lies within the bound (19) on $E[g(y)|A]$, regardless of how the imputed values are computed. Comparison of (22) and (1) shows that θ equals $E[g(y)|A]$ if these untestable conditions hold:

$$(24a)\quad P(A|z = 0) = P(x^* \in A|z = 0)$$

$$(24b)\quad E[g(y)|A, z = 0] = E[g(y^*)|x^* \in A, z = 0].$$

3.3. An Empirical Example

In Section 2.3, we used NLSY79 data to estimate the probability that a member of the surveyed population is employed in 1991. These data may be used in the same way to estimate the probability that a member of the population is unemployed or out of the labor force. Consider, however, the problem of inference on the official unemployment rate as measured in the United States by the Bureau of Labor Statistics. The official unemployment rate is the probability of unemployment within the subpopulation of persons who are in the labor force. When the 1991 employment status of an NLSY79 sample member is not reported, data are missing not only on that individual's unemployment outcome but also on his or her membership in the labor force. Thus, inference on the official unemployment rate poses a joint censoring problem.

Using the notation introduced in Section 2.3, the quantity of interest is $P[y = 1 | y \in \{1, 2\}]$ or, perhaps, $P[y = 1 | \text{BASE}, y \in \{1, 2\}]$. Using the data in Table 1, we find that the empirical unemployment rate among the individuals who respond to the 1991 employment-status question and who report that they are in the labor force is $P[y = 1 | y \in \{1, 2\}, z = 1] = 297/4629 = 0.064$. To compute the effective response probability $P_e[z = 1 | y \in \{1, 2\}]$, we need the unconditional response probability $P(z = 1)$ and the probability $P[y \in \{1, 2\} | z = 1]$ of being in the labor force conditional on response. These are $P(z = 1) = 5556/6812 = 0.816$ and $P[y \in \{1, 2\} | z = 1] = (4332 + 297)/5556 = 0.833$. Hence $P_e[z = 1 | y \in \{1, 2\}] = 0.787$. Equation (19) now yields the following bound on the official unemployment rate:

$$(0.064)(0.787) \leq P[y = 1 | y \in \{1, 2\}] \leq (0.064)(0.787) + 0.213$$

or $0.050 \leq P[y = 1 | y \in \{1, 2\}] \leq 0.263$. The analogous computations conditioning on the event BASE yield $0.057 \leq P[y = 1 | \text{BASE}, y \in \{1, 2\}] \leq 0.164$. Using the NLSY79 sampling weights yields 0.055 as the weighted estimate of the official unemployment rate.

As in the previous example, we estimate asymptotic joint Bonferroni confidence intervals for the estimates of the bounds. The derivations in this example are more lengthy than in the previous one, and so are given in an Appendix. The results are

$$0.044 \leq \text{lower bound on } P[y = 1 \mid y \in \{1, 2)] \leq 0.057$$

$$0.251 \leq \text{upper bound on } P[y = 1 \mid y \in \{1, 2)] \leq 0.277.$$

and

$$0.050 \leq \text{lower bound on } P[y = 1 \mid \text{BASE}, y \in \{1, 2)] \leq 0.065$$

$$0.153 \leq \text{upper bound on } P[y = 1 \mid \text{BASE}, y \in \{1, 2)] \leq 0.178.$$

As in the previous example, these confidence intervals are much narrower than the widths of the estimated bounds. Here, as before, identification is the dominant inferential problem.

4. Regressor Censoring

In this section, we assume that only observations of x are missing. Realizations of y are always observed. Thus z = 1 indicates that x is observed and z = 0 that x is missing. Regressor censoring, like joint censoring, has not previously been studied. The analysis below shows that regressor censoring generates its own distinctive identification problem.

4.1. Identification and the Asymptotic Bias of Estimates

To determine the identifiability of $E[g(y) \mid A]$, we combine the decomposition (1) with Bayes Theorem (16) to write

$$(25)\quad E[g(y)|A] = E[g(y)|A, z = 1]\,\frac{P(A|z = 1)P(z = 1)}{P(A|z = 1)P(z = 1) + P(A|z = 0)P(z = 0)}$$

$$+\; E[g(y)|A, z = 0]\,\frac{P(A|z = 0)P(z = 0)}{P(A|z = 1)P(z = 1) + P(A|z = 0)P(z = 0)}.$$

When y and x are jointly censored, the available sample data reveal nothing about $E[g(y)|A, z = 0]$ and $P(A|z = 0)$. When only x is censored, however, the sample data reveal the distribution $P(y|z = 0)$. Knowledge of $P(y|z = 0)$ implies restrictions on $E[g(y)|A, z = 0]$, and so we are able to narrow the bound obtained in Section 3.

To begin, observe that

$$(26)\quad P(y|z = 0) = P(y|A, z = 0)\cdot p + P(y|\bar{A}, z = 0)\cdot(1 - p),$$

where $p \equiv P(A|z = 0)$ and where $\bar{A}$ is the complement of A. Let Ψ denote the set of all distributions on Y. Proposition 1 of Horowitz and Manski (1995) shows that, if p is known, (26) implies this sharp restriction on $P(y|A, z = 0)$:

$$(27)\quad P(y|A, z = 0) \in \Psi(p) \equiv \Psi \cap \{[P(y|z = 0) - (1 - p)\psi]/p,\ \psi \in \Psi\}.$$

Hence we have this sharp p-dependent restriction on $E[g(y)|A, z = 0]$:

$$(28)\quad E[g(y)\,|\,A,\ z = 0] \ \in\ G(p) \ \equiv\ [\int g(y)d\psi,\ \ \psi \in \Psi(p)].$$

Let $g_0(p) \equiv \inf [h\colon h \in G(p)]$ and $g_1(p) \equiv \sup [h\colon h \in G(p)]$ be the sharp lower and upper bounds on the value of $E[g(y)\,|\,A, z = 0]$. Then (25) implies this sharp p-dependent bound on $E[g(y)\,|\,A]$:

$$(29)\quad E[g(y)\,|\,A,\ z = 1]\ \frac{P(A\,|\,z = 1)P(z = 1)}{P(A\,|\,z = 1)P(z = 1) + p\cdot P(z = 0)}$$

$$+\ \ g_0(p)\ \frac{p\cdot P(z = 0)}{P(A\,|\,z = 1)P(z = 1) + p\cdot P(z = 0)}$$

$$\leq\ \ E[g(y)\,|\,A]\ \ \leq$$

$$E[g(y)\,|\,A,\ z = 1]\ \frac{P(A\,|\,z = 1)P(z = 1)}{P(A\,|\,z = 1)P(z = 1) + p\cdot P(z = 0)}$$

$$+\ \ g_1(p)\ \frac{p\cdot P(z = 0)}{P(A\,|\,z = 1)P(z = 1) + p\cdot P(z = 0)}\ .$$

Horowitz and Manski (1995) give explicit expressions for $g_0(p)$ and $g_1(p)$ when $g(\cdot)$ is the identity function $g(y) = y$ and when $g(\cdot)$ is the indicator function $g(y) = 1[y \in B]$, where B is a subset of Y.

Of course p is not known, so the bound (29) is not computable. The

available sharp bound on $E[g(y)|A]$ is obtained from (29) by minimizing the lower bound over $p \in [0, 1]$ and by maximizing the upper bound over p. So we have this sharp bound on $E[g(y)|A]$:

$$(30)\quad \inf_p \{ E[g(y)|A, z = 1] \frac{P(A|z = 1)P(z = 1)}{P(A|z = 1)P(z = 1) + p \cdot P(z = 0)} + g_0(p) \frac{p \cdot P(z = 0)}{P(A|z = 1)P(z = 1) + p \cdot P(z = 0)} \}$$

$$\leq E[g(y)|A] \leq$$

$$\sup_p \{ E[g(y)|A, z = 1] \frac{P(A|z = 1)P(z = 1)}{P(A|z = 1)P(z = 1) + p \cdot P(z = 0)} + g_1(p) \frac{p \cdot P(z = 0)}{P(A|z = 1)P(z = 1) + p \cdot P(z = 0)} \} .$$

The sharp bound on the bias of an estimate θ follows immediately.

Regressor censoring yields all of the sample data available under joint censoring of outcomes and regressors, so the present bound must be a subset of the bound previously obtained under joint censoring. To verify this, observe that $K_0 \leq g_0(p) \leq g_1(p) \leq K_1$ for all p. Hence, (30) is a subset of the bound

$$(31)\quad \inf_{p} \{ E[g(y)|A, z = 1] \frac{P(A|z = 1)P(z = 1)}{P(A|z = 1)P(z = 1) + p\cdot P(z = 0)} + K_0 \frac{p\cdot P(z = 0)}{P(A|z = 1)P(z = 1) + p\cdot P(z = 0)} \}$$

$$\leq E[g(y)|A] \leq$$

$$\sup_{p} \{ E[g(y)|A, z = 1] \frac{P(A|z = 1)P(z = 1)}{P(A|z = 1)P(z = 1) + p\cdot P(z = 0)} + K_1 \frac{p\cdot P(z = 0)}{P(A|z = 1)P(z = 1) + p\cdot P(z = 0)} \} .$$

The solution to (31) is the bound (19) obtained under joint censoring.

The bound under regressor censoring typically is a proper subset of the bound under joint censoring. In fact, $E[g(y)|A]$ may even be identified in the presence of regressor censoring. To show this, we consider the extreme case in which the observable distribution $P(y|z = 0)$ is degenerate, with all its mass at some value $c \in Y$. Then $g_0(p) = g_1(p) = g(c)$ for all $p > 0$. The bound (30) reduces to

$$(32a)\quad E[g(y)\,|\,A,\ z = 1] \ \le\ E[g(y)\,|\,A] \ \le\ E[g(y)\,|\,A,\ z = 1]\cdot P_e(z = 1\,|\,A) + g(c)\cdot P_e(z = 0\,|\,A)$$

when $E[g(y)\,|\,A,\ z = 1] \le g(c)$ and

$$(32b)\quad E[g(y)\,|\,A,\ z = 1]\cdot P_e(z = 1\,|\,A) + g(c)\cdot P_e(z = 0\,|\,A) \ \le\ E[g(y)\,|\,A] \ \le\ E[g(y)\,|\,A,\ z = 1]$$

when $E[g(y)\,|\,A,\ z = 1] \ge g(c)$. Here $P_e(z = 1\,|\,A)$ is the effective response probability defined in (18).

Analyzing joint censoring of outcomes and regressors, we found that the bound on $E[g(y)\,|\,A]$ widens toward the noninformative interval $[K_0,\ K_1]$ as the effective response probability falls towards zero. Examining (32), we find that the bound on $E[g(y)\,|\,A]$ widens as $P_e(z = 1\,|\,A)$ falls, but its limit is the informative interval whose endpoints are $E[g(y)\,|\,A,\ z = 1]$ and $g(c)$. If $P(y\,|\,z = 0)$ happens to have all its mass at a point c such that $g(c) = E[g(y)\,|\,A,\ z = 1]$, then $E[g(y)\,|\,A]$ is identified.

4.2. Estimation Using Weights and Imputations

WEIGHTS: When regressors are censored, the weighted average estimate (4) continues to be computable and its probability limit remains

$E[s(w)\cdot g(y)\,|\,A,\ z = 1]$. As in the case of joint censoring of outcomes and regressors, weights of the form (7) are generally not computable except for the unweighted estimate $E[g(y)\,|\,A,\ z = 1]$. The unweighted estimate lies within the bound (30), and so is a logically possible value of $E[g(y)\,|\,A]$.

IMPUTATION: When regressors are censored, each sample member who does not report x may be assigned an imputed value x^*. The quantity $E[g(y)\,|\,A]$ may then be estimated by the sample average

$$(33)\quad \theta_N = \frac{1}{N_{1A} + N^*_{0A}} \sum_{i \in N(1,\,A)} g(y_i) + \sum_{i \in N^*(0,\,A)} g(y_i)$$

$$= \pi_N \frac{1}{N_{1A}} \sum_{i \in N(1,\,A)} g(y_i) + (1 - \pi_N) \frac{1}{N^*_{0A}} \sum_{i \in N^*(0,\,A)} g(y_i)\ ,$$

where $N(1, A)$, $N^*(0, A)$, and π_N were defined previously. The probability limit of the estimate is

$$(34)\quad \theta \equiv E[g(y)\,|\,A,\ z = 1]\cdot\pi + E[g(y)\,|\,x^* \in A,\ z = 0]\cdot(1 - \pi),$$

where π was defined in (23).

Comparison of (34) and (1) shows that θ equals $E[g(y)\,|\,A]$ if these two untestable conditions hold:

(35a) $P(A \mid z = 0) = P(x^* \in A \mid z = 0)$

(35b) $E[g(y) \mid A, z = 0] = E[g(y) \mid x^* \in A, z = 0]$.

θ always lies within the bound (30) on $E[g(y) \mid A]$, regardless of how the imputed values are computed. To see this, set $p = P(x^* \in A \mid z = 0)$ in (29). Then the p-dependent bound on $E[g(y) \mid A]$ is

(36) $E[g(y) \mid A, z = 1]\cdot\pi + g_0(\pi)\cdot(1 - \pi) \leq E[g(y) \mid A] \leq E[g(y) \mid A, z = 1]\cdot\pi + g_1(\pi)\cdot(1 - \pi)$.

The quantity $E[g(y) \mid x^* \in A, z = 0]$ necessarily lies in the interval $[g_0(p), g_1(p)]$. Hence θ lies within the p-dependent bound (36) and, a fortiori, within the available bound (30).

5. Total and Partial Joint Censoring

It is easy enough to think of situations in which researchers face some mixture of the three polar missing data problems analyzed in Sections 2 through 4. If anything, such mixtures are the norm in empirical work. There are many distinct cases that warrant attention, and it is impossible to treat them all here. We shall, however, examine one empirically important type of mixed

censoring -- some respondents are missing data on the outcome and all regressors, while others are missing data on the outcome and a subset of the regressors.

Mixtures of total and partial joint censoring may arise in many ways, but are especially prevalent in longitudinal surveys. Failure to interview sample members in the base year generates joint censoring of outcomes and all regressors. Attrition of respondents who are interviewed in the base year generates censoring of outcomes and a subset of regressors. We shall use this application to motivate the analysis below. We examine the identification problem in Section 5.1 and give an empirical example in Section 5.2.

5.1. Identification

To formalize the problem, we define two response indicators, z_1 and z_2. Let $z_1 = 1$ if a sample member is interviewed in the base year, at which time the regressors x_1 are measured. Let $z_2 = 1$ if a respondent is interviewed at a specified subsequent year, at which time the outcome y and additional regressors x_2 are measured. Assume that only those sample members interviewed in the base year are eligible to be interviewed thereafter. Hence $z_2 = 1 \Rightarrow z_1 = 1$.

We want to learn $E[g(y) \mid A_1, A_2] \equiv E[g(y) \mid x_1 \in A_1, x_2 \in A_2]$. To study its identifiability, we begin with the decomposition

$$(37)\quad E[g(y)|A_1, A_2] = E[g(y)|A_1, A_2, z_2 = 1]\cdot P(z_2 = 1|A_1, A_2) + E[g(y)|A_1, A_2, z_1 = 1, z_2 = 0]\cdot P(z_1 = 1, z_2 = 0|A_1, A_2) + E[g(y)|A_1, A_2, z_1 = 0]\cdot P(z_1 = 0|A_1, A_2).$$

The quantities $E[g(y)|A_1, A_2, z_1 = 1, z_2 = 0]$ and $E[g(y)|A_1, A_2, z_1 = 0]$ are not identified and can have any values between K_0 and K_1. Also,

$$(38)\quad P(z_1 = 1, z_2 = 0|A_1, A_2) + P(z_1 = 0|A_1, A_2) = 1 - P(z_1 = 1, z_2 = 1|A_1, A_2) = 1 - P(z_2 = 1|A_1, A_2).$$

Therefore,

$$(39)\quad E[g(y)|A_1, A_2, z_2 = 1]\cdot P(z_2 = 1|A_1, A_2) + K_0\cdot[1 - P(z_2 = 1|A_1, A_2)] \le E[g(y)|A_1, A_2] \le E[g(y)|A_1, A_2, z_2 = 1]\cdot P(z_2 = 1|A_1, A_2) + K_1\cdot[1 - P(z_2 = 1|A_1, A_2)].$$

The quantity $P(z_2 = 1|A_1, A_2)$ is not identified, but a tight lower bound can be found. By Bayes Theorem,

$$(40)\quad P(z_2 = 1|A_1, A_2) = \frac{P(A_1, A_2|z_2 = 1)P(z_2 = 1)}{P(A_1, A_2|z_2 = 1)P(z_2 = 1) + P(A_1, A_2|z_2 = 0)P(z_2 = 0)}.$$

Moreover,

$$(41)\quad P(A_1, A_2|z_2 = 0) = P(A_1, A_2|z_1 = 1, z_2 = 0)\, P(z_1 = 1|z_2 = 0) + P(A_1, A_2|z_1 = 0, z_2 = 0)\, P(z_1 = 0|z_2 = 0).$$

Nothing is known about $P(A_1, A_2|z_1 = 0, z_2 = 0)$ except that it is between 0 and 1. The quantity $P(A_1, A_2|z_1 = 1, z_2 = 0)$ is not identified, but this quantity is bounded from above by $P(A_1|z_1 = 1, z_2 = 0)$, which is identified. Therefore,

$$(42)\quad P(z_2 = 1|A_1, A_2) \geq P(A_1, A_2|z_2 = 1)P(z_2 = 1)/D,$$

where

$$(43)\quad D \equiv P(A_1, A_2|z_2 = 1)P(z_2 = 1) + P(A_1|z_1 = 1, z_2 = 0)P(z_1 = 1, z_2 = 0) + P(z_1 = 0).$$

Hence we obtain the same bound as under joint censoring of outcomes and regressors (see (19)), except that the effective response probability is now

$$(44)\quad P_e(z_2 = 1|A_1, A_2) = P(A_1, A_2|z_2 = 1)P(z_2 = 1)/D.$$

5.2. An Empirical Example

A simple modification of the example considered in Section 3.3 yields an illustration of mixed total and partial joint censoring. In Section 3.3, we wanted to learn the official unemployment rate in the entire population in the year 1991. Now let the objective be to learn the official unemployment rate in the subpopulation of white males. The NLSY79 data reveals the race and sex of every respondent who is interviewed in the base year, but data on employment status are available only for those individuals who are interviewed in 1991 and who report their employment status then. Thus, the event A_1 = {white male} and A_2 = {in the labor force in 1991}.

Table 2: EMPLOYMENT STATUS OF WHITE MALE NLSY79 RESPONDENTS

Employment Status	Number of Respondents
Employed (y = 2)	1900
Unemployed (y = 1)	114
Out of Labor Force (y = 0)	170
Ever-interviewed Nonrespondents	255
Total Ever-Interviewed White Males	2439

The NLSY79 employment-status data for ever-interviewed white males are presented in Table 2. These data show that the empirical unemployment rate among

white males is $P(y = 1|A_1, A_2, z_2 = 1) = 114/(1900 + 114) = 0.0566$. The data in Tables 1 and 2 imply, by (42), that the effective response rate is $P_e(z_2 = 1|A_1, A_2) = 0.6781$. Hence the implied bound on the official unemployment rate of white males is $0.038 \leq P(y = 1|A_1, A_2) \leq 0.360$. The estimate using the NLSY79 sampling weights is 0.054. The present bounds are wider than those found in Section 3.3 for the population unemployment rate because the additional conditioning here leads to a lower effective response rate.

6. The Identifying Power of Increasing Response Rates

In the administration of surveys, it may be possible to increase response rates through various means. Interview response rates may be increased by efforts to find sample members who initially cannot be located or to obtain consent from sample members who initially refuse to be interviewed. Item response rates may be increased through more intensive interviewing of respondents or, in some cases, by finding other sources for the missing data. The findings of Sections 2 through 5 imply that even relatively small levels of nonresponse can seriously degrade identification. Therefore, increasing the response rate may substantially improve identifiability. We characterize these improvements in this section, focusing on the cases of outcome censoring and joint censoring. The cases of regressor censoring and mixed censoring, which are

more complex, are not considered here.

In what follows, we suppose that a survey is administered in two stages, the second stage being an effort to increase response rates relative to the first stage. Let $z_a = 1$ if a sample member provides full information on (y, x) in the first stage; $z_a = 0$ otherwise. The second stage elicits responses only when $z_a = 0$. Let $z_b = 1$ if $z_a = 1$ or if the sample member responds fully in the second stage; $z_b = 0$ otherwise.

6.1. Outcome Censoring

In the case of outcome censoring, (2) implies that the sharp bounds on $E[g(y)|A]$ after the first and second stages are

$$(45)\quad E[g(y)|A, z_a = 1]\cdot\pi_a(A) + K_0\cdot[1 - \pi_a(A)] \le E[g(y)|A] \le E[g(y)|A, z_a = 1]\cdot\pi_a(A) + K_1\cdot[1 - \pi_a(A)]$$

and

$$(46)\quad E[g(y)|A, z_b = 1]\cdot\pi_b(A) + K_0\cdot[1 - \pi_b(A)] \le E[g(y)|A] \le E[g(y)|A, z_b = 1]\cdot\pi_b(A) + K_1\cdot[1 - \pi_b(A)],$$

where $\pi_a(A) \equiv P(z_a = 1|A)$ and $\pi_b(A) \equiv P(z_b = 1|A)$. Comparison of (45) and (46)

shows that increasing the response rate on y conditional on A from $\pi_a(A)$ to $\pi_b(A)$ decreases the width of the bound on $E[g(y)|A]$ from $(K_1 - K_0)[1 - \pi_a(A)]$ to $(K_1 - K_0)[1 - \pi_b(A)]$. It can be shown that the second-stage bound is not just narrower than the first-stage one; it is a subset thereof.

6.2. Joint Censoring of Outcomes and Regressors

In the case of joint censoring of outcomes and regressors, the first-stage and second-stage bounds have the same forms as (45) and (46) except that the unidentified response probabilities $\pi_a(A)$ and $\pi_b(A)$ are replaced by the effective response probabilities

$$(47a) \quad \pi_{ae}(A) \equiv \frac{P(A|z_a = 1)\pi_a}{P(A|z_a = 1)\pi_a + (1 - \pi_a)} = \frac{P(A, z_a = 1)}{P(A, z_a = 1) + (1 - \pi_a)}$$

and

$$(47b) \quad \pi_{be}(A) \equiv \frac{P(A|z_b = 1)\pi_b}{P(A|z_b = 1)\pi_b + (1 - \pi_b)} = \frac{P(A, z_b = 1)}{P(A, z_b = 1) + (1 - \pi_b)},$$

where $\pi_a \equiv P(z_a = 1)$ and $\pi_b \equiv P(z_b = 1)$ are the marginal response rates after the first and second stages.

The relationship between $\pi_{ae}(A)$ and $\pi_{be}(A)$ is not as transparent as the

relationship between $\pi_a(A)$ and $\pi_b(A)$, which is

$$(48)\quad \pi_b(A) = \pi_a(A) + P(z_a = 0,\ z_b = 1 \mid A).$$

To see how $\pi_{ae}(A)$ and $\pi_{be}(A)$ are related, observe that

$$\begin{aligned}(49)\quad P(A,\ z_b = 1) &= P(A,\ z_a = 1) + P(A,\ z_a = 0,\ z_b = 1)\\ &= P(A,\ z_a = 1) + P(A \mid z_a = 0,\ z_b = 1)P(z_a = 0,\ z_b = 1)\\ &= P(A,\ z_a = 1) + P(A \mid z_a = 0,\ z_b = 1)(\pi_b - \pi_a).\end{aligned}$$

Before the second-stage interviews are carried out, the value of $P(A \mid z_a = 0, z_b = 1)$ is unknown. Letting $P(A \mid z_a = 0, z_b = 1)$ vary over its logical range [0,1], (49) implies this sharp bound on $P(A,\ z_b = 1)$:

$$(50)\quad P(A,\ z_a = 1) \le P(A,\ z_b = 1) \le P(A,\ z_a = 1) + (\pi_b - \pi_a).$$

Hence $\pi_{be}(A)$ lies in the interval

$$(51)\quad \frac{P(A,\ z_a = 1)}{P(A,\ z_a = 1) + (1 - \pi_b)} \le \pi_{be}(A) \le \frac{P(A,\ z_a = 1) + (\pi_b - \pi_a)}{P(A,\ z_a = 1) + (1 - \pi_a)}.$$

The lower bound on $\pi_{be}(A)$ is larger than the value of $\pi_{ae}(A)$ given in (47a); hence

increasing the marginal response rate from π_a to π_b necessarily improves the identifiability of $E[g(y)|A]$. The width of the bound decreases from $(K_1 - K_0)[1 - \pi_{ae}(A)]$ to $(K_1 - K_0)[1 - \pi_{be}(A)]$. The magnitude of the reduction in width is illustrated in the next section.

6.3. An Empirical Example

To illustrate the findings of Section 6.2, let us return to the example presented in Section 3.3. There we considered the problem of inference on the official unemployment rate in the entire population. The estimated bound was $0.050 \leq P(y = 1|A) \leq 0.263$, where $A = [y \in \{1, 2\}]$. The estimated asymptotic Bonferroni confidence intervals for this estimate were

$$0.044 \leq \text{lower bound on } P(y = 1|A) \leq 0.057$$

$$0.251 \leq \text{upper bound on } P(y = 1|A) \leq 0.277.$$

Thus, the width of the estimated bound is 0.213. The narrowest width within the Bonferroni region is 0.194.

Now suppose that an additional stage of interviewing reduces the number of nonrespondents by half, from 1256 to 628. Then the data in Table 1 imply that

$$\pi_a(A) = 5556/6812 = 0.816$$

$$\pi_b(A) = 6184/6812 = 0.908$$

$$\pi_b(A) - \pi_a(A) = 628/6812 = 0.092$$

$$P(y \in A, z_a = 1) = (4332 + 297)/6812 = 0.680$$

$\pi_{ae}(A) = (0.680)/(0.680 + 0.184) = 0.787.$

The lower and upper bounds on $\pi_{be}(A)$ are

$\pi_{beL}(A) = (0.680)/(0.680 + 0.092) = 0.881$

$\pi_{beU}(A) = (0.680 + .092)/(0.680 + 0.184) = 0.893.$

Using $\pi_{beL}(A)$ as the effective response rate, which corresponds to setting $P(A \mid z_a = 0, z_b = 1) = 0$, yields the widest identification bounds at the reduced nonresponse rate. The resulting estimate of the bound on $P(y = 1 \mid A)$ is $0.056 \leq P(y = 1 \mid A) \leq 0.175$.

The width of this bound is 0.119, whereas the width of the bound at the first-stage nonresponse rate was 0.213. It is revealing to view this improvement from the perspective of classical statistical inference. The first-stage widths of the Bonferroni confidence intervals on the lower and upper bounds are 0.013 and 0.026. Drawing an unlimited number of additional respondents at random from the population would decrease the widths of these already quite narrow confidence intervals to zero, but would yield no improvement in the identifiability of the official unemployment rate. In contrast, converting just 628 nonrespondents into respondents yields a major improvement in our ability to infer the unemployment rate.

7. Conclusion

Survey nonresponse makes identification of population statistics problematic. It does not, however, preclude identification of bounds on population statistics. This paper has shown how identified bounds on unidentified population statistics can be obtained under several forms of nonresponse.

The bounds have been illustrated with empirical examples. These have demonstrated that nonresponse can make identification the dominant problem in inference, with sampling variation a second-order concern. The numerical examples have illustrated that efforts to increase response rates can have large benefits in terms of identification.

In applied research, it is customary to achieve identification in the presence of survey nonresponse through the use of weights, imputations or, more generally, models of the nonresponse process that place a priori restrictions on the relevant distributions in the nonresponding subpopulation. These restrictions usually are not testable. Moreover, corresponding to each point within the bounds identified by the data alone, there is a model that uniquely identifies that point. Thus, there are infinitely many mutually exclusive models that are consistent with the identifiable features of the data and the data cannot be used to determine which, if any, of the models is correct. This does not necessarily mean that models should not be used, but it is important to

understand the ambiguity that is inherent in any effort to use a model to achieve identification in the presence of nonresponse. The results presented in this paper are aimed at promoting such an understanding.

Appendix: Asymptotic Bonferroni Confidence Intervals Under Joint Censoring

Consider $P[y = 1 | y \in \{1, 2\}]$. The computations conditioning on the event BASE are analogous.

Some algebra shows that the lower bound given in (19) may be rewritten as $P_{11}/(1 - P_{01})$, where $P_{11} \equiv P(y = 1, z = 1)$ and $P_{01} \equiv P(y = 0, z = 1)$. Let $\hat{}$ signify estimators of population quantities. A Taylor series expansion yields

$$\frac{\hat{P}_{11}}{1 - \hat{P}_{01}} - \frac{P_{11}}{1 - P_{01}} =$$

$$[P_{11}/(1 - P_{01})][(\hat{P}_{11} - P_{11})/P_{11} + (\hat{P}_{01} - P_{01})/(1 - P_{01})] + o_p(N^{-1/2}).$$

Hence the variance of the limiting distribution of the lower bound is

$$V_{LB} = [(P_{11}/(1 - P_{01})]^2 [(1 - P_{11})/P_{11} - P_{01}/(1 - P_{01})].$$

The upper bound given in (19) may be rewritten as $(P_{11} + P_{\cdot 0})/(1 - P_{01})$, where $P_{\cdot 0} \equiv P(z = 0)$. A Taylor series expansion yields

$$(\hat{P}_{11} + \hat{P}_{\cdot 0})/(1 - \hat{P}_{01}) - (P_{11} + P_{\cdot 0})/(1 - P_{01}) =$$

$$[(P_{11} + P_{\cdot 0})/(1 - P_{01})][(\delta_{11} + \delta_{\cdot 0})/(P_{11} + P_{\cdot 0}) + \delta_{01}/(1 - P_{01})] + o_p(N^{-1/2}),$$

where $\delta_{ij} = \hat{P}_{ij} - P_{ij}$. Hence the variance of the limiting distribution is

$$\begin{aligned} V_{UB} = {} & [(P_{11} + P_{\cdot 0})/(1 - P_{01})]^2 \\ & \cdot\{[P_{11}(1 - P_{11}) + P_{\cdot 0}(1 - P_{\cdot 0}) - 2P_{11}P_{\cdot 0}]/(P_{11} + P_{\cdot 0})^2 \\ & + P_{01}(1 - P_{01})/(1 - P_{01})^2 - 2P_{01}(P_{11} + P_{\cdot 0})/[(P_{11} + P_{\cdot 0})(1 - P_{01})]\}. \end{aligned}$$

V_{LB} and V_{UB} can be estimated consistently by substituting estimators for P_{11}, P_{01}, and $P_{\cdot 0}$. This yields the estimated asymptotic Bonferroni 95 percent confidence regions reported in the text.

References

Center for Human Resource Research (1992) "NLS Handbook 1992. The National Longitudinal Surveys of Labor Market Experience," Columbus, Ohio: The Ohio State University.

Cochran, W., F. Mosteller, and J. Tukey (1954), Statistical Problems of the Kinsey Report on Sexual Behavior in the Human Male, Washington, D.C.: American Statistical Association.

Frankel, M., H. McWilliams, and B. Spencer (1983), "National Longitudinal Survey of Labor Force Behavior, Youth Survey: Technical Sampling Report," National Opinion Research Center, Chicago.

Horowitz, J. and C. Manski (1995), "Identification and Robustness with Contaminated and Corrupted Data," Econometrica, 63, 281-302.

MaCurdy, T., M. Gritz, and T. Mroz (1994), "An Evaluation of the NLSY," Department of Economics, Stanford University, Stanford, California.

Manski, C. (1994), "The Selection Problem," in C. Sims (editor) Advances in Econometrics: Sixth World Congress, Cambridge, England: Cambridge University Press.

Manski, C. (1995), Identification Problems in the Social Sciences, Cambridge, Mass.: Harvard University Press.

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