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

March 1994

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 SES-9223220.

Abstract

Censoring of outcomes (selection) is a common consequence of survey nonresponse and attrition in panels, and has received much attention. Joint censoring of regressors and outcomes is also common, but it has remained unexplored. This paper shows that the problem of identification when regressors and outcomes are jointly censored is equivalent to the problem of identification that would arise if only outcomes were censored but the censoring rate were higher than the actual one. As in the selection problem, one can identify informative bounds on conditional probability distributions, though the distributions themselves are not identified. These findings are illustrated using CPS income data and PSID data on welfare transitions. The identifying power of increasing the response rate to a survey is investigated.

1. Selection and Joint Censoring

Each year the U.S. Bureau of the Census uses data from household interviews in the March Current Population Survey (CPS) to estimate the distribution of income in the United States during the previous year. For example, U.S. Bureau of the Census (1993, Table 2, page 5) reports that in the year 1992, the fraction of all households with incomes below $15,000 was .24. Disaggregating by type of household, the Census Bureau reports that .55 of all households were married-couple families and the fraction of married-couple family households with incomes below $15,000 was .10. The Census Bureau reports that .12 of all households were female-headed families with no husband present and that the fraction of such households with incomes below $15,000 was .42.

Among the many considerations that must be kept in mind in interpreting these figures is nonresponse to the CPS. U.S. Bureau of the Census (1993, page C-10) states that 5 percent of the 60,000 households in the March 1993 CPS sample were not interviewed. How should the Census Bureau figures on income distribution be interpreted in light of this nonresponse? The complexity of the CPS sampling protocol and the Census Bureau practice of imputing missing income data makes this a difficult question to answer in a completely satisfactory manner. Imagine, however, that the CPS were a random sample of U.S. households and that accurate income responses were obtained from all sampled households who were

interviewed. We shall address the problem of nonresponse under these simplifying assumptions.

Consider first the Census Bureau report that the fraction of all households with incomes below $15,000 was .24. Under the maintained simplifying assumptions, this is a consistent estimate of the fraction of interviewable households with incomes below $15,000. Lacking information about the income distribution of non-interviewable households, we can only say that the fraction of such households with incomes below $15,000 lies between zero and one. The Census Bureau estimates that .95 of all households are interviewable and .05 are non-interviewable. These estimates imply that the fraction of all households with incomes below $15,000 is at least (.24 x .95) + (0 x .05) $\approx$ .23 and no more than (.24 x .95) + (1 x .05) $\approx$ .28. This illustrates the familiar *selection problem*. The conclusion that the fraction of all households lies in the interval [.23,.28] illustrates the worst-case bound studied in Manski (1989,1994). The essential point here is that because of nonresponse by some households to the CPS, the true fraction of households with incomes below $15,000 is not identified by the sampling process. The sampling process does, however, identify bounds on the fraction.

Now consider the Census Bureau report that the fraction of married-couple households with incomes below $15,000 was .10. Under the maintained assumption, this is a consistent estimate of

the fraction of interviewable married-couple households with incomes below $15,000. As before, the fraction of non-interviewable households with incomes below $15,000 lies between zero and one. The CPS data do not, however, reveal the fraction of married-couple households who are interviewable; the reason, of course, is that the household types of non-interviewed households are not known. So inference on the fraction of married-couple households with incomes below $15,000 poses a problem more serious than the selection problem. This is the problem of *joint censoring* of regressors and outcomes.

Let us now distinguish more abstractly between selection and joint censoring, using the problem of attrition in panel surveys to provide context. Suppose that a base survey administered to a random sample of a population provides observations on some covariates x_1 for all persons in the sample. Suppose that a followup survey is administered but that only some members of the sample respond. Let z = 1 for members of the original sample who respond to the followup and let z = 0 for those who do not. Thus, z = 0 indicates attrition from the sample. Suppose that the followup survey measures some covariates x_2 and an outcome y for those sample members with z = 1.

Suppose that one wishes to use the available data to learn about the conditional probability distribution $P(y \mid x_1 \in A_1, x_2 \in A_2)$. Here $A_1 \subset X_1$ and $A_2 \subset X_2$ are specified sets of values for x_1 and x_2,

while X_1 and X_2 are the domains of x_1 and x_2. If $A_2 = X_2$, only outcomes are censored and so one faces the selection problem. If A_2 is a proper subset of X_2, one faces the problem of inference on regressions when outcomes and regressors are jointly censored.

Inference in the presence of joint censoring, like inference in the presence of selection, is fundamentally a problem of identification. Joint censoring causes conditional probability distributions to be unidentified unless one is willing to make untestable assumptions about the censoring process. This paper investigates what is identified when one does not make such assumptions. The result turns out to be quite simple: the problem of identification when regressors and outcomes are jointly censored is equivalent to the problem of identification that would arise if only outcomes were censored (the selection problem) but the censoring rate were higher than the actual one. As in the selection problem, one can identify informative bounds on conditional probability distributions, though the distributions themselves are not identified.

Section 2 establishes this result. Section 3 illustrates the result using the CPS income data and PSID data on welfare transitions. Section 4 examines identification when censoring is known to be exogeneous in either of two senses. Section 5 investigates the identifying power of increasing the response rate to a survey. We find that, in some situations, small increases in

the response rate can produce large gains in identification. Section 6 gives conclusions.

2. Identification Under Joint Censoring

2.1. The Effective Selection Probability

We begin by deriving the *effective selection probability* that enables us to represent the problem of identification under joint censoring as a problem of identification when only outcomes are censored. Define $P(y|A_1,A_2) \equiv P(y|x_1 \in A_1, x_2 \in A_2)$ and write

$$(1) \qquad P(y|A_1,A_2) = P(y|A_1,A_2,z=1)P(z=1|A_1,A_2) + P(y|A_1,A_2,z=0)P(z=0|A_1,A_2).$$

If only y is censored, the sampling process identifies all of the quantities on the right side of (1) except for the distribution $P(y|A_1,A_2,z=0)$ of y conditional on censoring. If (y,x_2) are jointly censored, the sampling process no longer identifies the selection probability $P(z=1|A_1,A_2)$ but does still imply a sharp lower bound, this being

$$(2)\quad P_L(z=1|A_1,A_2) \equiv \frac{P(A_2|A_1,z=1)P(z=1|A_1)}{P(A_2|A_1,z=1)P(z=1|A_1) + P(z=0|A_1)}.$$

We call $P_L(z=1|A_1,A_2)$ the *effective selection probability*.

The fact that $P_L(z=1|A_1,A_2)$ is the sharp lower bound on $P(z=1|A_1,A_2)$ is easily shown by using Bayes Theorem to write

$$(3)\quad P(z=1|A_1,A_2) = \frac{P(A_2|A_1,z=1)P(z=1|A_1)}{P(A_2|A_1,z=1)P(z=1|A_1) + P(A_2|A_1,z=0)P(z=0|A_1)}.$$

Inspect the right side of (3). The sampling process identifies $P(A_2|A_1,z=1)$, $P(z=1|A_1)$, and $P(z=0|A_1)$, but is uninformative about $P(A_2|A_1,z=0)$. Setting $P(A_2|A_1,z=0) = 1$ yields the sharp lower bound on $P(z=1|A_1,A_2)$, and setting $P(A_2|A_1,z=0) = 0$ yields the sharp upper bound. The result is

$$(4)\quad P_L(z=1|A_1,A_2) \le P(z=1|A_1,A_2) \le 1.$$

Holding A_1 fixed, the effective selection probability decreases as the set A_2 shrinks. The limiting cases are

$$(5a)\quad P(A_2|A_1,z=1) = 1 \Rightarrow P_L(z=1|A_1,A_2) = P(z=1|A_1)$$

and

(5b) $P(A_2|A_1,z=1) = 0 \Rightarrow P_L(z=1|A_1,A_2) = 0.$

So inference on the conditional distribution $P(y|A_1,A_2)$ becomes more difficult as the conditioning event A_2 shrinks in probability, and is not feasible at all when A_2 has probability zero. In particular, the sampling process is uninformative about $P(y|A_1,x_2)$ if the random variable x_2 is continuous.

Note that joint censoring is more informative than the *truncated sampling* process drawing observations only from the subpopulation with $z = 1$. Truncated sampling identifies $P(y|A_1,A_2,z=1)$ but reveals nothing at all about $P(z=1|A_1,A_2)$. So the effective selection probability under truncated sampling always is zero.

2.2. Implications

Manski's (1989, 1994) analysis of the selection problem generalizes to the problem of joint censoring. One just replaces the actual selection probability $P(z=1|A_1,A_2)$ with the effective selection probability $P_L(z=1|A_1,A_2)$, and then analyzes the inferential problem as if only outcomes were censored.

The basic finding is a sharp bound on $E[g(y)|A_1,A_2]$ for any function $g(y)$. Write

$$(6)\quad E[g(y)\,|\,A_1,A_2] = E[g(y)\,|\,A_1,A_2,z=1]P(z=1\,|\,A_1,A_2) + E[g(y)\,|\,A_1,A_2,z=0]P(z=0\,|\,A_1,A_2).$$

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. If only y is censored, then all the quantities on the right side of (6) are identified except for $E[g(y)\,|\,A_1,A_2,z=0]$, which must lie between K_0 and K_1. Hence

$$(7)\quad K_0 + P(z=1\,|\,A_1,A_2)\{E[g(y)\,|\,A_1,A_2,z=1] - K_0\} \le E[g(y)\,|\,A_1,A_2] \le K_1 - P(z=1\,|\,A_1,A_2)\{K_1 - E[g(y)\,|\,A_1,A_2,z=1]\}$$

If x_2 is censored jointly with y, then $P(z=1\,|\,A_1,A_2)$ is not identified and so the bound (7) is not computable. The sampling process does, however, reveal that $P(z=1\,|\,A_1,A_2) \ge P_L(z=1\,|\,A_1,A_2)$. Replacing $P(z=1\,|\,A_1,A_2)$ with $P_L(z=1\,|\,A_1,A_2)$ in (7) yields the best computable bound on $E[g(y)\,|\,A_1,A_2]$, namely

$$(8)\quad K_0 + P_L(z=1\,|\,A_1,A_2)\{E[g(y)\,|\,A_1,A_2,z=1] - K_0\} \le E[g(y)\,|\,A_1,A_2] \le K_1 - P_L(z=1\,|\,A_1,A_2)\{K_1 - E[g(y)\,|\,A_1,A_2,z=1]\}.$$

So joint censoring of (y,x_2) with selection probability $P(z=1\,|\,A_1,A_2)$ implies the same bound on $E[g(y)\,|\,A_1,A_2]$ as does censoring of y alone with selection probability $P_L(z=1\,|\,A_1,A_2)$.

Just as the bound (7) forms the starting point for derivation of further results on the selection problem, (8) implies further results on joint censoring. Consider the probability $P(y \in B | A_1, A_2)$ that y lies in a specified set B. Letting $g(y) \equiv 1[y \in B]$ in (8) implies that

$$(9) \quad P_L(z=1|A_1,A_2)P(y\in B|A_1,A_2,z=1) \leq P(y\in B|A_1,A_2) \leq 1 - P_L(z=1|A_1,A_2)[1 - P(y\in B|A_1,A_2,z=1)].$$

If y is real-valued, (9) implies that for all $t \in R_1$,

$$(10) \quad P_L(z=1|A_1,A_2)P(y\leq t|A_1,A_2,z=1) \leq P(y\leq t|A_1,A_2) \leq 1 - P_L(z=1|A_1,A_2)[1 - P(y\leq t|A_1,A_2,z=1)].$$

Bound (10) can be inverted to obtain bounds on quantiles of $P(y|A_1,A_2)$. Let $\alpha \in (0,1)$, and let $q(\alpha,A_1,A_2)$ denote the α-quantile of y conditional on the event $(x_1 \in A_1, x_2 \in A_2)$. Let $r(\alpha,A_1,A_2)$ be the $[1-(1-\alpha)/P_L(z=1|A_1,A_2)]$-quantile of $P(y|A_1,A_2,z=1)$ if $P_L(z=1|A_1,A_2) > 1 - \alpha$ and to be $-\infty$ otherwise. Define $s(\alpha,A_1,A_2)$ to be the $[\alpha/P_L(z=1|A_1,A_2)]$-quantile of $P(y|A_1,A_2,z=1)$ if $P_L(z=1|A_1,A_2) \geq \alpha$ and to be ∞ otherwise. It can be shown that (see Manski, 1994)

(11) $r(\alpha, A_1, A_2) \leq q(\alpha, A_1, A_2) \leq s(\alpha, A_1, A_2)$.

3. Illustrations

3.1. Nonresponse in the CPS

Let us illustrate the bound (10) using the CPS data on incomes given in Section 1. Let y denote household income, x_1 denote household nationality, x_2 denote household type, and z indicate whether a household is interviewable. Let A_1 be the set of all American households and let A_2 be the set of households composed of married-couple families. In Section 1 we left unresolved the matter of inference on $P(y \leq 15,000 \mid A_1, A_2)$, but we now can answer the question. The Census Bureau reports .10, .55, and .95 as consistent estimates of $P(y \leq 15,000 \mid A_1, A_2, z=1)$, $P(A_2 \mid A_1, z=1)$, and $P(z=1 \mid A_1)$. It follows that .91 is a consistent estimate of the effective selection probability $P_L(z=1 \mid A_1, A_2)$, and that [.09,.18] is a consistent estimate of the sharp bound on $P(y \leq 15,000 \mid A_1, A_2)$.

Now modify A_2 to be the set of households composed of female-headed families with no husband present. Then .42 and .12 are the estimates of $P(y \leq 15,000 \mid A_1, A_2, z=1)$ and $P(A_2 \mid A_1, z=1)$. So the estimate of $P_L(z=1 \mid A_1, A_2)$ is .70 and the estimated bound on $P(y \leq 15,000 \mid A_1, A_2)$ is [.08,.38].

Observe how the width of the bound changes as we modify the conditioning event A_2. In Section 1, we found that the fraction of all American households with incomes below \$15,000 must lie in an interval of width .05, namely [.23,.28]. Here we have found that the fractions of married-couple households and female-headed households with incomes below \$15,000 must lie in intervals of width .09 and .30 respectively. This illustrates how inference becomes increasingly difficult as $P(A_2 | A_1, z=1)$ falls.

3.2. Attrition in the PSID

Panel data are often used to learn transition probabilities. Attrition is a common concern in the analysis of transitions. We illustrate using the Panel Study of Income Dynamics (PSID) to examine welfare transitions from one year to the next. See Hill (1992) for background on the PSID.

Let y_t indicate whether a woman receives welfare benefits in year t; in particular, let $y_t = 1$ if a woman receives Aid to Families with Dependent Children (AFDC), and $y_t = 0$ otherwise. Let v_t indicate a woman's family status in year t; in particular, let $v_t = 1$ if a woman has one or two children and is either a wife or head of household, and $v_t = 0$ otherwise. Let $z_t = 1$ if a woman is interviewable about welfare and family status in year t, and $z_t = 0$ otherwise.

We use the PSID to infer $P(y_{t+1}=1 \mid z_t=1, y_t=1, v_t=1, v_{t+1}=1)$ for the year t = 1987. In the notation used earlier, $y = y_{t+1}$, $z = z_{t+1}$, A_1 is the event $(z_t=1, y_t=1, v_t=1)$, and A_2 is the event $(v_{t+1}=1)$. That is, we restrict attention to those women who are interviewable about their status in 1987, who have one or two children in both 1987 and 1988, and who receive AFDC in 1987. Among this group, we seek to learn what fraction receive AFDC in 1988. The problem is that not all members of the group are interviewable about their 1988 status. For those who are not, we do not observe (y_{t+1}, v_{t+1}).

(In fact, this description of the problem does not quite match the structure of the PSID. Each year's interview determines AFDC participation in the previous year and family status at the time of the interview. To simplify our illustration, we treat the family status response as applying to the previous year. So we use the 1988 wave of the PSID to provide welfare and family data on 1987 and the 1989 wave to provide data on 1988.)

Let N(*) be the number of PSID respondents for whom condition * holds. Here are the relevant data:

$N(z_t=1, y_t=1, v_t=1) = 172$

$N(z_t=1, y_t=1, v_t=1, z_{t+1}=1) = 165$

$N(z_t=1, y_t=1, v_t=1, z_{t+1}=1, v_{t+1}=1) = 141$

$N(z_t=1, y_t=1, v_t=1, z_{t+1}=1, v_{t+1}=1, y_{t+1}=1) = 101.$

Using PSID sample frequencies to estimate corresponding probabilities yields these estimates:

$$P_N(z_{t+1}=1 \mid z_t=1, y_t=1, v_t=1) = 165/172$$

$$P_N(v_{t+1}=1 \mid z_t=1, y_t=1, v_t=1, z_{t+1}=1) = 141/165$$

$$P_N(y_{t+1}=1 \mid z_t=1, y_t=1, v_t=1, z_{t+1}=1, v_{t+1}=1) = 101/141.$$

Thus, whereas the estimate of the selection probability is 165/172, the estimate of the effective selection probability is, by (2),

$$P_{LN}(z_{t+1}=1 \mid z_t=1, y_t=1, v_t=1) = \frac{(141/165)(165/172)}{(141/165)(165/172) + 7/172} = 141/148.$$

Applying (9), the estimate of the lower bound on $P(y_{t+1}=1 \mid z_t=1, y_t=1, v_t=1, v_{t+1}=1)$ is $(141/148)(101/141) = 101/148$ and the estimate of the upper bound is $1 - (141/148)(1 - 101/141) = 108/148$.

4. Exogenous Selection

We have thus far presumed that the researcher possesses no information about the censoring process. It is often assumed that selection is exogenous in some sense. When such an assumption is

made, the bounds of Section 2 can be made narrower.

One sense of exogeneity is conveyed by assuming that the outcome y is statistically independent of z, conditional on the event $(x_1 \in A_1, x_2 \in A_2)$; that is,

$$(12) \quad P(y \mid A_1, A_2) = P(y \mid A_1, A_2, z=1) = P(y \mid A_1, A_2, z=0).$$

The distribution $P(y \mid A_1, A_2, z=1)$ is identified by the sampling process, so $P(y \mid A_1, A_2)$ is identified if (12) holds.

A different sense of exogeneity is conveyed by assuming that the event $(x_2 \in A_2)$ is statistically independent of z, conditional on the event $(x_1 \in A_1)$; that is,

$$(13) \quad P(A_2 \mid A_1) = P(A_2 \mid A_1, z=1) = P(A_2 \mid A_1, z=0).$$

It follows from (13) and (3) that

$$(14) \quad P(z=1 \mid A_1, A_2) = P(z=1 \mid A_1).$$

The right side of (14) is identified by the sampling process, so this exogeneity assumption makes the problem of inference when (y, x_2) are jointly censored equivalent to the problem of inference when only outcomes are censored.

5. The Identifying Power of Increasing Response Rates

In survey research, it is sometimes possible to increase the response rate through efforts to find and interview nonrespondents. Let $z_a = 1$ if a sampled person responds to a first interview attempt; $z_a = 0$ otherwise. Let $z_b = 1$ if $z_a = 1$ or if the person responds to a second interview attempt made in an effort to increase the response rate; $z_b = 0$ otherwise. From the perspective of identification, what is the value of the second interview attempt? The answer depends on the prior information about the censoring process possessed by the researcher.

Suppose that selection is known to be exogenous in the sense of (12). Then $P(y|A_1,A_2)$ is identified except in the limiting case of a zero response rate. So increasing the response rate has no identifying power as long as the response rate on the first interview attempt is positive. Increasing the response rate does, however, have value from the perspective of statistical inference: it increases the available sample size and so enables more precise estimation of $P(y|A_1,A_2)$.

Suppose that selection is known to be exogenous in the sense of (13), so (14) holds. Now increasing the response rate does have identifying power. To keep the discussion simple, consider the problem of inference on an event probability $P(y\in B|A_1,A_2)$. Define $\Pi_a \equiv P(z_a=1|A_1)$ and $\Pi_b \equiv P(z_b=1|A_1)$. Inspection of (9) shows that

increasing the response rate of persons in the set A_1 from Π_a to Π_b reduces the width of the bound on $P(y \in B \mid A_1, A_2)$ from $1 - \Pi_a$ to $1 - \Pi_b$.

Finally suppose that nothing is known about the censoring process. Now the initial and revised widths of the bound on $P(y \in B \mid A_1, A_2)$ are $1 - \Pi_{aL}$ and $1 - \Pi_{bL}$, where

$$(15a) \qquad \Pi_{aL} \equiv \frac{P(A_2 \mid A_1, z_a=1)\Pi_a}{P(A_2 \mid A_1, z_a=1)\Pi_a + (1 - \Pi_a)} \frac{P(A_2, z_a=1 \mid A_1)}{P(A_2, z_a=1 \mid A_1) + (1 - \Pi_a)} =$$

and

$$(15b) \qquad \Pi_{bL} \equiv \frac{P(A_2 \mid A_1, z_b=1)\Pi_b}{P(A_2 \mid A_1, z_b=1)\Pi_b + (1 - \Pi_b)} \frac{P(A_2, z_b=1 \mid A_1)}{P(A_2, z_b=1 \mid A_1) + (1 - \Pi_b)} =$$

are the effective selection probabilities after the first and second interview attempts.

To see how Π_{aL} and Π_{bL} are related, observe that

$$(16) \quad P(A_2, z_b=1 \mid A_1) = P(A_2, z_a=1 \mid A_1) + P(A_2, z_a=0, z_b=1 \mid A_1)$$

$$= P(A_2, z_a=1 \mid A_1) + P(A_2 \mid A_1, z_a=0, z_b=1)P(z_a=0, z_b=1 \mid A_1)$$

$$= P(A_2, z_a=1 \mid A_1) + P(A_2 \mid A_1, z_a=0, z_b=1)(\Pi_b - \Pi_a).$$

The quantity $P(A_2 | A_1, z_a=0, z_b=1)$ may take any value in the interval [0,1]. So we have this bound on $P(A_2, z_b=1 | A_1)$:

$$(17) \quad P(A_2, z_a=1 | A_1) \leq P(A_2, z_b=1 | A_1) \leq P(A_2, z_a=1 | A_1) + (\Pi_b - \Pi_a).$$

Hence Π_{bL} satisfies the bound

$$(18) \quad \frac{P(A_2, z_a=1 | A_1)}{P(A_2, z_a=1 | A_1) + (1 - \Pi_b)} \leq \Pi_{bL} \leq \frac{P(A_2, z_a=1 | A_1) + (\Pi_b - \Pi_a)}{P(A_2, z_a=1 | A_1) + (1 - \Pi_a)}.$$

The lower bound on Π_{bL} is larger than the value of Π_{aL}. This confirms that increasing the response rate has identifying power. The upper bound on Π_{bL} exceeds Π_{aL} by more than the increase $(\Pi_b - \Pi_a)$ in the response rate.

The practical significance of this result is illustrated in Table 1, which shows how increasing the response rate by .05 (i.e. making $\Pi_b = \Pi_a + .05$) can change the effective selection probability. For each of several values of the initial response rate Π_a and of the event probability $P(A_2 | A_1, z=1)$, the Table reports the initial effective selection probability Π_{aL} and the bounds (18) on the revised effective selection probability Π_{bL}. Perhaps the most interesting feature of the table is that the lower bound on Π_{bL} exceeds the value of Π_{aL} by more than .05 when Π_a equals 0.8 or 0.9. In these cases, we can unambiguously conclude that increasing the

survey response rate by .05 yields a larger increase in the effective selection probability.

TABLE 1

$P(A_2 \mid A_1, z=1)$	Π_a	Π_{aL}	Bound on Π_{bL}
0.3	0.6	0.3103	[0.3396, 0.3966]
	0.7	0.4118	[0.4565, 0.5098]
	0.8	0.5455	[0.6154, 0.6591]
	0.9	0.7297	[0.8437, 0.8649]
0.5	0.6	0.4286	[0.4615, 0.5000]
	0.7	0.5385	[0.5833, 0.6154]
	0.8	0.6667	[0.7273, 0.7500]
	0.9	0.8182	[0.9000, 0.9091]
0.7	0.6	0.5122	[0.5455, 0.5732]
	0.7	0.6203	[0.6622, 0.6835]
	0.8	0.7368	[0.7887, 0.8026]
	0.9	0.8630	[0.9265, 0.9315]

It is revealing to view the table entries from the perspective of classical statistical inference. Suppose that there were no censoring and that the sample included N persons for whom the event (A_1, A_2) occurred. Then an asymptotically valid 95% percent sampling confidence interval for $P(y \in B \mid A_1, A_2)$ would have width $2(1.96)\{P(y \in B \mid A_1, A_2)[1-P(y \in B \mid A_1, A_2)]/N\}^{1/2}$. The confidence interval is widest when $P(y \in B \mid A_1, A_2) = 0.5$, in which case its width is $2(1.96)(0.5)/N^{1/2}$. For example, the maximum width is 0.0620 when N = 1000 and is 0.0438 when N = 2000.

The widths of identification bounds and of 95% sampling confidence intervals for $P(y \in B \mid A_1, A_2)$ are not strictly comparable,

but each measures a form of uncertainty about $P(y \in B \mid A_1, A_2)$. The width of an identification bound measures uncertainty caused by survey nonresponse, and the width of a 95% confidence interval measures uncertainty caused by sampling error. Table 1 indicates that uncertainty due to nonresponse dominates uncertainty due to sampling error unless the sample is small or the response rate is close to 1. Increasing the response rate by 0.05 narrows the width of the identification bound on $P(y \in B \mid A_1, A_2)$ by at least 0.0263, and can narrow this width by as much as .1352. (The former number is the smallest difference in the table between the lower bound on Π_{bL} and the value of Π_{aL}, while the latter one is the largest difference between the upper bound on Π_{bL} and the value of Π_{aL}.) In contrast, doubling the sample size from 1000 to 2000 narrows the sampling confidence interval by at most 0.0620 – 0.0438 = 0.0182. In the context of the selection problem, similar examples showing that uncertainty due to nonresponse typically dominates uncertainty due to sampling error were presented forty years ago by Cochran, Mosteller, and Tukey (1954, pages 274 – 282).

6. Conclusion

Censoring of outcomes is a common problem in the analysis of survey data, and has received much attention. Joint censoring of regressors and outcomes is also a common problem, but has somehow remained invisible. In this paper, we have shown that the problem of identification when regressors and outcomes are jointly censored is equivalent to the problem of identification that would arise if only outcomes were censored (the selection problem) but the censoring rate were higher than the actual one. As in the selection problem, one can identify informative bounds on conditional probability distributions, though the distributions themselves are not identified. We hope that these simple findings will impel researchers facing joint censoring problems to properly recognize the implications for empirical analysis.

References

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.

Hill, M. (1992), The Panel Study of Income Dynamics: A User's Guide, Newbury Park, CA: Sage Publications.

Manski, C. (1989), "Anatomy of the Selection Problem," Journal of Human Resources, 24: 343-360.

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

U.S. Bureau of the Census (1993), Current Population Reports, Series P-60, No. 184, Money Income of Households, Families, and Persons in the United States: 1992, Washington, D.C.: U.S. Government Printing Office.