CREDIBLE INTERVAL ESTIMATES FOR OFFICIAL STATISTICS WITH SURVEY NONRESPONSE

Charles F. Manski
Department of Economics and Institute for Policy Research
Northwestern University

First Public Draft: February 2013
This Draft: June 2014

Abstract

Government agencies commonly report official statistics based on survey data as point estimates, without accompanying measures of error. Agencies could measure sampling error using established statistical principles, but it is more challenging to measure nonsampling errors. This paper considers error due to survey nonresponse. The standard practice has been to use weights and imputations to implement assumptions that nonresponse is conditionally random. I review modern research deriving interval estimates that make no assumptions about the values of missing data. To demonstrate the implications for official statistics, I use data from the U. S. Current Population Survey to form interval estimates for median household income, the family poverty rate, and the unemployment rate. I then explore some of the middle ground between interval estimation making no assumptions and point estimation assuming that nonresponse is conditionally random.

This research was supported in part by National Science Foundation grant SES-1129475. I am grateful to Aanchal Jain for excellent research assistance and to Bruce Spencer for helpful discussions. I have benefitted from the opportunity to present aspects of this work in seminars at Carnegie Mellon University, the Federal Reserve Bank of New York, and the University of Essex. I have also benefitted from the comments of two anonymous reviewers.

1. Introduction

Government agencies commonly report official statistics based on survey data as point estimates, without accompanying measures of error. News releases present the estimates with little if any mention of the possibility of error. Technical publications documenting data and methods often acknowledge verbally that official statistics are subject to error, but they may provide only limited guidance on the magnitude of sampling errors and typically do not attempt to quantify nonsampling errors. Some prominent American examples include the reporting of employment and income statistics by the Bureau of Labor Statistics (BLS) and the Census Bureau.

BLS Reporting of Employment Statistics: On the first Friday of each month, the BLS issues *The Employment Situation*, a monthly news release reporting official employment statistics for the previous month. For example, the BLS reported this on October 5, 2012 (U. S. Bureau of Labor Statistics, 2012): "The unemployment rate decreased to 7.8 percent in September, and total nonfarm payroll employment rose by 114,000." The unemployment-rate statistic is based on data on households sampled in the Current Population Survey (CPS). The one on nonfarm employment is based on data collected from employer establishments sampled in the Current Employment Statistics survey (CES).

The BLS monthly news release reports employment statistics as point estimates, without measures of potential error. A Technical Note issued with the news release contains a section on *Reliability of the estimates* that acknowledges the possible presence of errors, beginning with the statement "Statistics based on the household and establishment surveys are subject to both sampling and nonsampling error." The section describes the conventional use of standard errors and confidence intervals to measure sampling error, providing some numerical illustrations.

The Technical Note then turns to nonsampling errors, stating that they "can occur for many reasons, including the failure to sample a segment of the population, inability to obtain information for all respondents

in the sample, inability or unwillingness of respondents to provide correct information on a timely basis, mistakes made by respondents, and errors made in the collection or processing of the data." The Note does not indicate the magnitudes of the nonsampling errors that may be present in the employment statistics. □

Census Reporting of Income Statistics: Each year the Census Bureau reports statistics on the income distribution based on data collected in the Annual Social and Economic (ASEC) Supplement to the CPS. In a news release issued September 12, 2012, the Census Bureau declared (U. S. Census Bureau, 2012A): "The nation's official poverty rate in 2011 was 15.0 percent, with 46.2 million people in poverty. After three consecutive years of increases, neither the poverty rate nor the number of people in poverty were statistically different from the 2010 estimates." Thus, the Census release provided point estimates, acknowledged but did not quantify sampling error, and did not mention nonsampling error.

The Census Bureau's P60 *Current Population Report* provides numerous statistics characterizing the income distribution and measures sampling error by providing 90-percent confidence intervals for various estimates (U. S. Census Bureau, 2012B). However, the report does not measure nonsampling errors. A supplementary technical document describes some sources of nonsampling error, but it does not quantify them (U. S. Census Bureau, 2012C). □

Reporting official statistics as point estimates manifests a common tendency of policy analysts to project *incredible certitude*, encouraging the public to believe that errors are small and inconsequential (Manski, 2013, Ch. 1). In the absence of agency guidance, persons who understand that official statistics are subject to error must fend for themselves and conjecture the error magnitudes. Thus, users of official statistics may misinterpret the information that the statistics provide.

Agencies could mitigate misinterpretation of official statistics if they were to measure and report potential errors. Using established statistical principles, agencies can report sampling error for levels and

temporal changes in important statistics based on survey data. It is more challenging to measure nonsampling errors because there are many sources of such errors and there has been no consensus about how to measure them. Yet these facts do not justify ignoring nonsampling error. Having agency analytical staffs make good-faith efforts to measure error would be more informative than reporting official statistics as if they are truths.

Aiming to advance discourse on practical ways to measure nonsampling error, this paper considers error due to survey nonresponse. Nonresponse is only one of the sources of nonsampling error that affect the validity of official statistics. I focus on it for three reasons.

First, nonresponse is common in the surveys used to compute important official statistics. Unit and item nonresponse regularly make key data missing for substantial fractions of the persons sampled. For example, consider the official statistics on household income based on the ASEC Supplement to the CPS. During the period 2002-2012, 7 to 9 percent of the sampled households yielded no income data due to unit nonresponse and 41 to 47 percent of the interviewed households yielded incomplete income data due to item nonresponse.

Second, statistical agencies use traditional but untenable assumptions—namely that nonresponse is random conditional on specified observed covariates—to form point estimates with nonresponse. These assumptions are implemented as weights for unit nonresponse and imputations for item nonresponse. In particular, the Census Bureau applies *hot-deck* imputations to the CPS and other surveys, describing the hot deck this way (U. S. Census Bureau, 2006, p. 9-2):

"This method assigns a missing value from a record with similar characteristics, which is the hot deck. Hot decks are defined by variables such as age, race, and sex. Other characteristics used in hot decks vary depending on the nature of the unanswered question. For instance, most labor force questions use age, race, sex, and occasionally another correlated labor force item such as full- or part-time status."

Thus the agency staff select a vector of covariates for which response is complete and computes the empirical

distribution of the outcome of interest among sample members who have this covariate value and who report their outcomes. An outcome is imputed to a sample member with missing data by drawing a realization at random from the available empirical distribution.

The CPS documentation of hot-deck imputation offers no evidence that the method yields an outcome distribution for missing data that is close to the actual distribution of such outcomes. However, another Census Bureau document describing the American Housing Survey is revealing. The document states (U. S. Census Bureau, 2011):

"Some people refuse the interview or do not know the answers. When the entire interview is missing, other similar interviews represent the missing ones For most missing answers, an answer from a similar household is copied. The Census Bureau does not know how close the imputed values are to the actual values."

Indeed, lack of knowledge of the closeness of imputed values to actual ones is common.

The third reason to focus on nonresponse is that econometric research has shown how to form interval estimates that make no assumptions about nonresponse; see Manski (1989, 1994, 2003) and Horowitz and Manski (1998, 2000) *inter alia*. This research has also shown how to form confidence intervals that jointly measure sampling error and potential nonresponse error (e.g., Horowitz and Manski, 2000; Imbens and Manski, 2004). Thus, we know how to report credible interval estimates for official statistics with survey nonresponse.

Section 2 reviews basic results and specific findings on interval estimation that make no assumptions about nonresponse. The central problem is identification rather than statistical inference from finite samples. That is, in the absence of assumptions about nonresponse, one can only obtain bounds on the population parameters that official statistics intend to measure. These bounds provide the basis for formation of interval estimates that may be computed using available survey data.

Applying these findings, Section 3 uses data from the ASEC and the monthly CPS to form interval

estimates for median household income, the family poverty rate, and the unemployment rate. I provide one set of estimates that take into account item nonresponse alone and another that recognizes unit response as well. The estimates show vividly that item nonresponse poses a huge problem for inference on the American income distribution, and that unit nonresponse exacerbates the problem. While item nonresponse is a relatively minor source of error for the unemployment rate, unit nonresponse is highly consequential.

Section 4 explores some of the middle ground between interval estimation making no assumptions on nonresponse and traditional point estimation assuming that nonresponse is conditionally random. Interval estimation making no assumptions has maximal credibility, but it may be excessively conservative if agency analysts have some understanding of the nature of nonresponse. Point estimation assuming random nonresponse has maximal precision but may have little credibility. The middle ground derives interval estimates based on assumptions that may include random nonresponse as one among various possibilities. I do not recommend adoption of a particular middle-ground assumption for reporting of official statistics. I only pose some alternatives that statistical agencies may want to consider.

Section 5 makes concluding comments on measurement of nonresponse error in official statistics. I urge statistical agencies to report statistics that quantify potential error openly.

2. Interval Estimation Without Assumptions on Nonresponse: A Review of Basic Findings

Many official statistics aim to measure parameters of a probability distribution characterizing a population. For example, the poverty rate is intended to be the fraction of persons or families with income below specified levels. The unemployment rate is defined to be the fraction of unemployed persons among those who are in the labor force. In each case the statistic is a parameter of a probability distribution $P(y|x)$, where y is a specified outcome taking values in a space Y and x is a specified covariate taking values in a

space X . The parameter of interest, say $\theta[P(y|x)]$, may be the mean or a quantile of y within the sub-population with covariate value x .

Surveys such as the CPS draw stratified random samples of population units, ask sample members to report their values of (y, x) , and use the responses to estimate official statistics. Some sample members provide complete information and some choose not to respond to a subset of the questions posed (item nonresponse). Some provide no information at all (unit nonresponse), either because they choose not to be interviewed or because survey staff are unable to contact them. Thus, data on either or both of y and x may be missing for some sample members.

Estimation of population parameters with no assumptions about nonresponse is basically a matter of contemplating all the values that the missing data might take. Section 2.1 explains the problem in abstraction. Section 2.2 summarizes findings on outcome nonresponse, which is relatively simple to study.

2.1. The Inferential Problem in Abstraction

2.1.1. Identification

To study inference with missing data, it is productive to first consider identification and then statistical inference. In identification analysis, one supposes that all population units are sample members. Hence, one knows the distributions of y and x for units who report them. These distributions are $P(y|z_y = 1)$, $P(x|z_x = 1)$, and $P(y, x|z_y = z_x = 1)$, where $z_y = 1$ (or 0) if a population unit would (or not) report y , and $z_x = 1$ (or 0) if the unit would (or not) report x . One also knows the distribution $P(z_y, z_x)$ of response.

Given knowledge of these observable distributions, one may determine what this implies about the parameter $\theta[P(y|x)]$. The generic finding is that $\theta[P(y|x)]$ lies in a set of values, say $H\{\theta[P(y|x)]\}$, called its *identification region* or *identified set*. The parameter is *point-identified* if the identification region contains just one point. It is *partially identified* if the region contains multiple values but is a proper subset of the

space of all logically possible values of the parameter (Manski, 2003).

The above suffices when considering identification of official statistics characterizing outcomes in a single time period (perhaps a month or year), using only the data collected in that period. For this purpose, it is not necessary to use notation explicitly marking the period of interest. Now suppose that data have been collected in periods $t = 1, \dots, T$ and suppose that one wants to use the data to learn the time-series of official statistics $\theta[P_t(y|x)]$, $t = 1, \dots, T$. Then the period-specific findings extend immediately to the time series. With no assumptions relating unknowns across time periods, the joint identification region for the time-series vector of statistics is the Cartesian product of the period-specific regions. That is,

$$(1) \quad H\{\theta[P_t(y|x)], t = 1, \dots, T\} = \prod_{t=1, \dots, T} H_t\{\theta[P_t(y|x)]\},$$

where $H_t\{\theta[P_t(y|x)]\}$ is the identification region for $\theta[P_t(y|x)]$ obtained using only the data collected in period t and the maintained period-specific assumptions.

Equation (1) provides the basis for determination of the identification region for any function of the time-series vector of statistics. For example, a common concern is to learn the change $\theta[P_t(y|x)] - \theta[P_{t-1}(y|x)]$ in an official statistic between two adjacent periods $t - 1$ and t . The Cartesian product form of the joint identification region for $\{\theta[P_{t-1}(y|x)], \theta[P_t(y|x)]\}$ implies that the sharp lower (upper) bound on the temporal change in the statistic is the lower (upper) bound of $H_t\{\theta[P_t(y|x)]\}$ minus the upper (lower) bound of $H_{t-1}\{\theta[P_{t-1}(y|x)]\}$.

2.1.2. Statistical Inference

With identification determined, one may study statistical inference when a sampling process draws a finite number of population units. Suppose for simplicity that one draws a random sample of N units. A natural way to estimate the parameter is to use the empirical distributions $P_N(y|z_y = 1)$, $P_N(x|z_x = 1)$, $P_N(y, x|z_y$

$= z_x = 1$), and $P_N(z_y, z_x)$ to estimate their population counterparts. This yields an estimate $H_N\{\theta[P(y|x)]\}$ of the identification region of the parameter. Given ordinary regularity conditions, the Strong Law of Large Numbers implies that this set-valued estimate is consistent; that is, it converges almost surely to the set $H\{\theta[P(y|x)]\}$ as $N \rightarrow \infty$.

One can also form confidence sets to measure the uncertainty created by sampling variation. Recall the standard definition of a confidence set for a real parameter θ . One first specifies a *coverage probability* α , where $0 < \alpha < 1$. One next considers alternative ways to use the sample data to form sets on the real line. Let $C(\cdot)$ be a set-valued function that maps the data into a set on the real line; thus, for each possible value ψ of the sample data, $C(\psi)$ is the set that results when the data are ψ . Then $C(\cdot)$ gives an α -confidence set for θ if $\text{Prob}[\psi: \theta \in C(\psi)] = \alpha$.

Although the statistics literature has focused on parameters that are point-identified, the standard definition of a confidence set also applies to parameters that are partially identified. In addition, one can define confidence sets for identification regions. Let $H(\theta)$ be the identification region for θ . Then $C(\cdot)$ gives an α -confidence set for $H(\theta)$ if $\text{Prob}[\psi: H(\theta) \subset C(\psi)] = \alpha$. An α -confidence set for $H(\theta)$ necessarily covers θ with probability at least α . This holds because the true value of θ lies in $H(\theta)$; hence, $\text{Prob}[\psi: \theta \in C(\psi)] \geq \text{Prob}[\psi: H(\theta) \subset C(\psi)]$. See Imbens and Manski (2004).

2.2. Outcome Nonresponse

Derivation of identification regions is particularly simple when only the outcome variable y is subject to nonresponse, the conditioning covariates x always being observed, and the parameter of interest respects stochastic dominance. I summarize findings for this case here, drawing on Manski (1989, 1994, 2003). See Horowitz and Manski (1998, 2000) for analysis of identification with covariate nonresponse. See Stoye (2010) for analysis of spread parameters, which do not respect stochastic dominance.

With nonresponse on outcomes alone, the structure of the inferential problem is displayed by the Law of Total Probability

$$(2) \quad P(y|x) = P(y|x, z = 1)P(z = 1|x) + P(y|x, z = 0)P(z = 0|x),$$

where $z \equiv z_y$. Supposing that all population units are sample members, the empirical evidence reveals $P(z|x)$. It also reveals the distribution $P(y|x, z = 1)$ of observable outcomes when $P(z = 1|x) > 0$. The evidence is uninformative regarding the distribution $P(y|x, z = 0)$ of missing outcomes, which may be any probability distribution on Y . Hence, the evidence reveals that $P(y|x)$ lies in the identification region

$$(3) \quad H[P(y|x)] = [P(y|x, z = 1)P(z = 1|x) + \gamma P(z = 0|x), \gamma \in \Gamma_Y],$$

where Γ_Y denotes the set of all probability distributions on the set Y .

The above concerns identification of the entire outcome distribution. Now consider identification of a parameter $\theta[P(y|x)]$. The identification region is the set of all the values it can take when $P(y|x)$ ranges over all of its feasible values. Thus, $H\{\theta[P(y|x)]\} = \{\theta(\eta), \eta \in H[P(y|x)]\}$.

2.2.1. Event Probabilities

The above derivation is straightforward but abstract. To begin to show its practical implications, first suppose that one wants to learn the probability $P(y \in B|x)$ that y falls in a specified set B . By the Law of Total Probability,

$$(4) \quad P(y \in B|x) = P(y \in B|x, z = 1)P(z = 1|x) + P(y \in B|x, z = 0)P(z = 0|x).$$

The empirical evidence reveals $P(y \in B|x, z = 1)$, $P(z = 1|x)$, and $P(z = 0|x)$, but provides no information on $P(y \in B|x, z = 0)$. The last quantity necessarily lies between zero and one. This yields the following sharp bound on $P(y \in B|x)$, reported in Manski (1989):

$$(5) \quad P(y \in B|x, z = 1)P(z = 1|x) \leq P(y \in B|x) \leq P(y \in B|x, z = 1)P(z = 1|x) + P(z = 0|x).$$

2.2.2. The Distribution Function and Quantiles

The bound on event probabilities has many applications. An immediate one is the bound that it implies on the distribution function of a real-valued outcome. Suppose that y is real-valued. The distribution function for $P(y|x)$ is $P(y \leq t|x)$, $t \in \mathbb{R}$. Application of (5) to $P(y \leq t|x)$ gives

$$(6) \quad P(y \leq t|x, z = 1)P(z = 1|x) \leq P(y \leq t|x) \leq P(y \leq t|x, z = 1)P(z = 1|x) + P(z = 0|x).$$

The feasible distribution functions are all increasing functions of t that take values between the lower and upper bounds in (6) for all values of t .

Bound (6) can be inverted to obtain sharp bounds on quantiles of $P(y|x)$. Consider the α -quantile $Q_\alpha(y|x)$. Let y_0 and y_1 denote the smallest and largest logically possible values of y ; thus, $y_0 \equiv \min_{y \in Y}$ and $y_1 \equiv \max_{y \in Y}$. Manski (1994) shows that the sharp lower and upper bounds on $Q_\alpha(y|x)$ are $r(\alpha, x)$ and $s(\alpha, x)$, where

$$(7) \quad r(\alpha, x) \equiv [\alpha - P(z = 0|x)]/P(z = 1|x) \text{ quantile of } P(y|x, z = 1) \text{ if } P(z = 0|x) < \alpha, \\ \equiv y_0 \text{ otherwise.}$$

$$(8) \quad s(\alpha, x) \equiv \alpha/P(z = 1|x) \text{ quantile of } P(y|x, z = 1) \text{ if } P(z = 0|x) \leq 1 - \alpha, \\ \equiv y_1 \text{ otherwise.}$$

2.2.3. Means of Functions of the Outcome

Finally, let $g(\cdot)$ be a specified function of y and consider inference on the conditional mean $E[g(y)|x]$. Suppose first that $g(\cdot)$ has bounded range, with sharp lower and upper bounds g_0 and g_1 . The Law of Iterated Expectations gives

$$(9) \quad E[g(y)|x] = E[g(y)|x, z = 1]P(z = 1|x) + E[g(y)|x, z = 0]P(z = 0|x).$$

The empirical evidence reveals $E[g(y)|x, z = 1]$ and $P(z|x)$. However, the evidence reveals nothing about $E[g(y)|x, z = 0]$, which can take any value in the interval $[g_0, g_1]$. Hence, the identification region for $E[g(y)|x]$ is the interval

$$(10) \quad H\{E[g(y)|x]\} \\ = [E[g(y)|x, z = 1]P(z = 1|x) + g_0P(z = 0|x), E[g(y)|x, z = 1]P(z = 1|x) + g_1P(z = 0|x)].$$

The situation changes if $g(\cdot)$ is unbounded from below or above; that is, if $g_0 = -\infty$ or $g_1 = \infty$. In such cases, result (10) still holds but has different implications whenever $P(z = 0|x) > 0$. The lower bound on $E[g(y)|x]$ is $-\infty$ if $g_0 = -\infty$ and is ∞ if $g_1 = \infty$. The identification region has infinite width but remains informative if $g(\cdot)$ is bounded from at least one side. However, it is the entire real line if $g(\cdot)$ is unbounded from both below and above. Thus, the presence of missing data makes assumptions a prerequisite for inference on the mean of an unbounded random variable.

2.2.4. Estimation of the Bounds

The bounds on quantiles and means may be obtained by the same simple argument, which holds for any parameter that respects stochastic dominance (see Manski, 2003, Sec. 1.3). Wherever data are missing,

insert the values of y that yield the smallest and largest values of the parameter of interest to obtain the lower and upper bounds. Given this, estimation of the bounds with sample data is easy.

To estimate the lower bound, one supposes that $y_i = y_0$ for each sample member i with missing data. One then computes the usual point estimate of the parameter. To estimate the upper bound, one likewise supposes that $y_i = y_1$ whenever observation i is missing. Thus, estimation of the bound simply requires two extreme imputations of each case of missing data.

3. Illustrative Interval Estimates: Median Household Income, Family Poverty, and Unemployment

Official statistics describing income and employment in the United States are based on data collected in the Current Population Survey. The basic CPS sampling unit is a household, which may include one or more families or unrelated individuals who reside together. Data on annual income of household members in the preceding calendar year are collected in the ASEC Supplement, which is administered in the period February through April. Data on the current employment status of civilian adult household members are collected in the monthly administration of the CPS.

Applying the findings of Section 2, I use ASEC data collected in 2002-2012 to form interval estimates of median household income and the fraction of families with income below the official poverty threshold in the years 2001-2011. I use monthly CPS data to form interval estimates of the unemployment rate in March of 2012-2012.

To keep the application simple, I do not seasonally adjust the unemployment rate, as does the BLS in its official monthly estimate. I do not weight the raw data to reflect geographic clustering in the CPS sample design. I also do not report sampling confidence intervals for the interval estimates. The ASEC and regular CPS sample sizes are so large that sampling uncertainty is a minor consideration relative to the

identification problem stemming from lack of knowledge of missing data.

4.1. Estimates Recognizing Item Nonresponse

The interval estimates in Table 1 show the potential implications of item nonresponse alone, ignoring unit nonresponse. The bound on the unemployment rate is easy to compute. The CPS questionnaire asks whether each adult member of the household is employed, looking for work, or out of the labor force. Data are missing when a response to this question is not obtained. The lower bound on the unemployment rate is the value that the rate would take if all persons with missing data were employed. The upper bound is the value that the rate would take if all persons with missing data were looking for work. See Horowitz and Manski (1998) for the proof.

The bounds on median household income and the family poverty rate would be similarly simple if the ASEC questionnaire were to pose one question asking for total household income and another asking for the total income of each family sub-unit of the household. Then the findings of Section 2.2 would apply immediately. In the absence of assumptions, the lower (upper) bound on median household income would be the value that the median would take if all households with missing data were to have zero (infinite) income. The lower (upper) bound on the family poverty rate would be the value that the rate would take if all families with missing data were to have infinite (zero) income.

In actuality, the ASEC questionnaire inquires about eighteen separate income components, ranging from earnings and pensions to dividends and public assistance. To determine total household income, the Census Bureau sums the responses obtained for these income components across the members of the household, imputing the values of missing data. The ASEC data shows a wide range of item nonresponse patterns, as households differ in the data they provide about the various income components.

For example, nonresponse to the question asking for earnings on the primary job ranges from 17 to

19 percent over the ten-year period. Nonresponse to the question on interest income ranges from 23 to 27 percent over the period. Nonresponse on dividend income ranges from 9 to 12 percent.

Whenever there is missing data on an income component, I take the lower bound on this component to be zero. Hence, the lower bound on total household or family income is the sum of the relevant reported income components. There is no similarly obvious finite upper bound on the value of a missing income component. One could set the upper bound as infinity, but this seems excessively cautious. The interval estimates in Table 1 take the upper bound on each missing income component to be the highest value reported by any member of the ASEC sample. Hence, the upper bound on total household or family income is the sum of the relevant income components, using reported values when available and the maximum value occurring in the sample when data are missing.

Examination of the table shows bounds that vary in their informativeness, which depends both on the fraction of cases with missing data and on the statistic under consideration. First consider median household income and the family poverty rate. The fraction of interviewed households with data missing on some income components is huge. In every year at least 0.41 of households and 0.38 of families had some missing income data. The bounds on the statistics are consequently very wide. For example, the bound on median household income in 2001 is [\$32000, \$102000] and the bound on the family poverty rate is [0.11, 0.32]. The bounds for 2011 are [\$38000, \$100000] and [0.14, 0.34] respectively. Thus, in the absence of assumptions on the missing data, the CPS data reveal little about median income or the poverty rate.

Now consider the unemployment rate. Among all civilian adults who are interviewed, the fraction with missing employment data is relatively small but increases over time, ranging from 0.0028 in 2002 to 0.0076 in 2012. The bounds on the unemployment rate correspondingly widen over time, ranging from [0.057, 0.062] in March 2002 to [0.078, 0.090] in March 2012.

Table 1 also shows point estimates that use Census Bureau imputations of missing data. The point estimates always lie within the bounds. This is necessarily the case given that the imputations are logically

possible values of the missing data.

[TABLE 1 HERE]

3.2. Estimates Recognizing Item and Unit Nonresponse

The above bounds, as wide as they are, only recognize item nonresponse. Consideration of unit nonresponse widens the bounds further. During the period 2002-2012, the fraction of sampled ASEC households that were not interviewed ranged from 0.069 to 0.089. These unit nonresponse rates are small relative to other major household surveys, yet they still are consequential in the absence of assumptions on the missing data. Unit nonresponse turns out to be particularly consequential for inference on the unemployment rate.

Table 2 reports interval estimates that recognize both item and unit nonresponse. Whereas Table 1A took sample size to be the number of interviewed households, Table 2A takes sample size to be the larger number of households at which an interview was attempted. Sample size excludes housing units that Census interviewers found to be vacant or demolished.

Table 2 does not include a column with point estimates. The Census Bureau does not impute data missing from unit nonresponse. Instead, the Bureau weights the interviewed households with weights derived from an assumption that unit nonresponse is random.

Interval estimation of median household income is straightforward. Each case of unit nonresponse yields exactly one missing household whose income is entirely unknown. Interval estimation of the family poverty rate and the unemployment rate runs into the conceptual difficulty that we do not know the composition of non-interviewed households. To simplify the application, I assume that every non-interviewed household contains exactly one family and one civilian adult in the labor force. Then each case of unit nonresponse yields exactly one missing family and one potentially unemployed person. To the extent

that non-interviewed households contain multiple families or multiple individuals in the labor force, the intervals reported in Table 2 would widen further.

To summarize the table, consider the years 2001 and 2011. Taking unit nonresponse into account, the bound on median household income in 2001 is [\$28200, \$123643] and the bound on the family poverty rate is [0.10, 0.36]. The bounds for 2011 are [\$32132, \$100000] and [0.13, 0.39] respectively. The bounds on the unemployment rate range from [0.053, 0.123] in March 2002 to [0.071, 0.162] in March 2012.

[TABLE 2 HERE]

4. Interval Estimation with Assumptions on Nonresponse

Interval estimates of official statistics that place no assumptions on the values of missing data are credible, easy to understand, and simple to compute. One might therefore think that it would be standard practice for government statistical agencies to report them. I am, however, unaware of any official statistics reported this way.

In personal communications, colleagues have sometimes remarked that interval estimates obtained without assumptions on nonresponse are not reported because they are "too wide to be informative." Indeed, the illustrative empirical analysis of Section 3 shows that, with current nonresponse rates, the intervals tend to be rather wide. Nevertheless, I would argue that the Census Bureau and the BLS should report them.

Even when wide, interval estimates obtained without assumptions on nonresponse are valuable for two reasons. First, they are maximally credible in the sense that they express all possible values of the statistic of interest, whatever values the missing data may take. Second, they make explicit the fundamental role that assumptions play in inferential methods that yield tighter findings. Wide bounds reflect real uncertainties that cannot be washed away by assumptions lacking credibility.

The above argument does not imply, of course, that statistical agencies should entirely refrain from making assumptions about nonresponse. Interval estimates obtained with no assumptions may be excessively conservative if agency analysts have some understanding of the nature of nonresponse. There is a vast middle ground between interval estimation with no assumptions and the standard practice of point estimation assuming that nonresponse is conditionally random. The middle ground obtains interval estimates based on assumptions that may include random nonresponse as one possibility. Manski (2003, Chapters 1 and 2); 2007, Chapters 3 and 4) characterize the identifying power of various such assumptions, including ones that use instrumental variables and that place shape restrictions on $P(y|x)$.

It is unlikely that any one middle-ground assumption will be appropriate in all settings. Hence, I do not propose adoption of any particular assumption for reporting of official statistics. Instead, this section suggests some alternatives that statistical agencies may want to consider. Given the historical prominence of assumptions that nonresponse is random, agencies may find it appealing to consider assumptions formalizing the conjecture that missing data are not too different from observed data. Section 4.1 poses assumptions that directly constrain the distribution of missing data. Section 4.2 considers ones that constrain the response propensities of persons with different outcomes. Section 4.3 discusses ones that relate nonresponse at different points in time. For simplicity, I restrict attention to outcome nonresponse and suppose that the outcome takes finitely many values.

4.1. Assumptions on the Distribution of Missing Outcomes

Assumptions that directly constrain the distribution of missing outcomes place $P(y|x, z = 0)$ within some set of outcome distributions. Abstractly, one might assume that $P(y|x, z = 0) \in \Gamma_{cY}$, where Γ_{cY} is a specified set of outcome distributions. Recall that, using the empirical evidence alone, the Law of Total Probability (2) yields identification region (3) for $P(y|x)$. The evidence combined with the assumption that

$P(y|x, z = 0) \in \Gamma_{cY}$ analogously yields the identification region (Manski, 2003, eq. 1.3)

$$(11) \quad H_c[P(y|x)] \equiv [P(y|x, z = 1)P(z = 1|x) + \gamma P(z = 0|x), \gamma \in \Gamma_{cY}].$$

Implementation of this abstract idea depends on how one specifies Γ_{cY} . A conjecture that missing data do not differ too much from observed data can be formalized by asserting that $P(y|x, z = 0)$ lies in some neighborhood of $P(y|x, z = 1)$. The specific assumption then depends on how one defines neighborhoods.

Various definitions of neighborhoods have been proposed in research on *sensitivity analysis*, which explores how inferential conclusions weaken as the maximum distance between $P(y|x, z = 0)$ and $P(y|x, z = 1)$ grows. Much work on sensitivity analysis has been developed for application to observational studies of treatment response, where outcomes of counterfactual treatments are never observed and attrition from the study may prevent observation of the outcomes of some realized treatments. See Robins, Rotnitzky, and Sharfstein (1999), Rosenbaum (2002), and the references cited therein.

Kline and Santos (2013) propose a form of sensitivity analysis for application to survey nonresponse. They define neighborhoods by the Kolmogorov-Smirnov distance between $P(y|x, z = 0)$ and $P(y|x, z = 1)$, considered uniformly across the covariate space. As a leading example, they consider the mixture model in which some fraction of nonresponse is random and the remaining fraction arises from an unknown mechanism. They relate this model to the Horowitz and Manski (1995) study of partial identification with corrupted sampling, where the available data are a mixture of actual observations of y and erroneous observations drawn from an unknown distribution, the source of each observation being unobservable. Manski (2007, p. 97) observes that the identification problem with corrupted sampling is more severe than that with survey nonresponse.

It is important to understand that the way one defines neighborhoods matters. To illustrate, I pose two possibilities, embodying alternative perspectives on how missing data may differ from observed

outcomes. First consider the mixture model mentioned above. This model asserts that

$$(12) \quad P(y|x, z=0) = (1 - \delta)P(y|x, z=1) + \delta\gamma,$$

where γ is an unknown outcome distribution and δ is the fraction of nonresponse drawn from γ . Then $\Gamma_{cY} = [(1 - \delta)P(y|x, z=1) + \delta\gamma, \gamma \in \Gamma_Y]$. It follows that

$$(13) \quad H_c[P(y|x)] = \{P(y|x, z=1)[P(z=1|x) + (1 - \delta)P(z=0|x)] + \gamma[\delta P(z=0|x)], \gamma \in \Gamma_Y\}.$$

Identification region (13) has the same form as the region (3) that would be obtained without assumptions, if the fraction of missing data were $\delta P(z=0|x)$ rather than $P(z=0|x)$. A version of this simple result appears in Kline and Santos (2013, eq. 8).

Alternatively, one might think it credible to assume that the probability with which each missing outcome value occurs is not too different from the corresponding probability for observed outcomes. Let $0 \leq \lambda_k \leq 1$ be the assumed maximum deviation between the probability of the event $[y = k]$ conditional on the outcome being missing and observed. The assumption asserts that

$$(14) \quad |P(y = k|x, z=0) - P(y = k|x, z=1)| \leq \lambda_k, \quad \text{all } k \in Y.$$

Hence,

$$(15) \quad \Gamma_{cY} = [\gamma \in \Gamma_Y: |\gamma(k) - P(y = k|x, z=1)| \leq \lambda_k, \quad \text{all } k \in Y].$$

4.2. Assumptions on Response Propensities Conditional on Outcomes

When outcome data are missing, the evidence reveals the response propensities $P(z = 1|x)$ that condition on covariates but only partially reveals the propensities $P(z = 1|x, y)$ that condition on covariates and outcomes. Assumptions that constrain the latter propensities may have identifying power for $P(y|x)$.

The tools used to exploit such assumptions are Bayes Theorem and the fact that probabilities of mutually exclusive and exhaustive events sum to one. Bayes Theorem gives

$$(16) \quad P(y = k|x) = P(y = k|x, z = 1)P(z = 1|x)/P(z = 1|x, y = k), \quad k \in Y.$$

For each k , the evidence reveals $P(y = k|x, z = 1)$ and $P(z = 1|x)$. Hence, assumptions on $P(z = 1|x, y = k)$ imply restrictions on $P(y = k|x)$. In particular, the maximum (minimum) feasible value of $P(z = 1|x, y = k)$ gives the minimum (maximum) feasible value of $P(y = k|x)$.

To begin, $P(z = 1|x, y = k)$ is a probability, so $0 \leq P(z = 1|x, y = k) \leq 1$. Summing $P(y = k|x)$ across $k \in Y$ gives a constraint on the vector $[P(z = 1|x, y = k), k \in Y]$ that uses the empirical evidence alone, namely

$$(17) \quad 1 = \sum_{k \in Y} P(y = k|x) = \sum_{k \in Y} P(y = k|x, z = 1)P(z = 1|x)/P(z = 1|x, y = k).$$

Further constraints may be posed as assumptions asserting that $[P(z = 1|x, y = k), k \in Y]$ lies in some set of $|Y|$ -dimensional vectors, say F . Combining the fact that $[P(z = 1|x, y = k), k \in Y]$ is a vector of probabilities that solves (17) and the assumption that the vector lies in F yields restrictions on $[P(y = k|x), k \in Y]$ via (16).

For example, one might think it credible to assume that each component of $[P(z = 1|x, y = k), k \in Y]$ lies in some neighborhood of $P(z = 1|x)$. That is, one might assume that

$$(18) \quad \alpha_k P(z = 1|x) \leq P(z = 1|x, y = k) \leq \beta_k P(z = 1|x), \quad k \in K$$

for specified constants $0 \leq \alpha_k \leq 1 \leq \beta_k$, $k \in Y$. Then $F = \times_{k \in Y} [\alpha_k P(z = 1|x), \beta_k P(z = 1|x)]$. The assumption of random nonresponse is the limit case where $\alpha_k = 1 = \beta_k$, $k \in Y$. In this case, application of (16) yields the familiar result $P(y = k|x) = P(y = k|x, z = 1)$, $k \in Y$.

4.2.1. Comparison of Assumptions Constraining Missing Outcomes and Response Propensities

It is illuminating to compare the present class of assumptions with that considered in Section 4.1. Both classes have identifying power. However, they may differ in at least three important respects: refutability, tractability, and transparency.

Assumptions that directly constrain the distribution of missing outcomes are nonrefutable—after all, the available data imply no constraints on the missing data. In contrast, some assumptions that constrain response propensities are refutable. That is, one may find that there exists no vector $[P(z = 1|x, y = k), k \in Y]$ that solves (17) and lies in F . If so, one should conclude that the vector does not lie in F . Note, however, that assumption (18) is nonrefutable, as it weakens the assumption of random nonresponse.

The assumptions also differ in the complexity of the identification regions that they imply. The region implied by assumption (12) has the same simple form as the one obtained without assumptions. The regions implied by the other assumptions are more complex. To implement assumptions on response propensities, one first finds the set of vectors $[P(z = 1|x, y = k), k \in Y]$ that solve (17) and lies in F . One then uses (16) to determine the implications for $P(y|x)$. The implied set of feasible values for $P(y|x)$ typically has no simple explicit form.

Considering tractability alone, one might think it convenient to pose assumptions of form (12). However, considerations of transparency may generate a countervailing force. Given that nonresponse is a behavioral phenomenon, wherein persons choose to be interviewed and to respond to the questions asked,

an analyst may find it natural to pose conjectures about how response behavior varies with personal characteristics. Such conjectures may be formalized as assumptions on response propensities. Contrariwise, an analyst may find it artificial to pose conjectures about the distribution of missing data, which is determined jointly by response behavior and by the population distribution of (y, x) . In particular, one may find it difficult to motivate the mixture model expressed in (12).

4.3. Assumptions Restricting Temporal Variation in Nonresponse

The assumptions considered above place restrictions on nonresponse in a specified time period. Assumptions may also relate nonresponse across time periods. Then the joint identification region for the time-series vector of a statistic may be a proper subset of the region obtained without these assumptions. The joint region generally has no simple explicit form, but it can be determined numerically.

Many temporal assumptions may be reasonable to conjecture. Among them, ones supposing that unknown quantities do not change too rapidly with time may be particularly credible when considering statistics on employment, income, and other quantities that are thought to vary relatively smoothly with time. There are multiple ways to formalize the idea. One might restrict the time-series variation of the period-specific distributions of missing outcome data $P_t(y|x, z = 0)$, $t = 1, \dots, T$. One might restrict the time-series variation of the period-specific response propensities $P_t(z = 1|x, y)$, $t = 1, \dots, T$. Or one might restrict the time-series variation of the period-specific population outcome distributions $P_t(y|x)$, $t = 1, \dots, T$.

Each type of assumption yields constraints on the time-series vector of official statistics, the specifics depending on how one formalizes the notion of restricting time-series variation. Moreover, one may combine temporal-variation assumptions with period-specific assumptions of the forms studied in Section 4.1 and 4.2 to achieve additional identifying power.

5. Conclusion

The prevailing practice in the reporting of official statistics is to acknowledge nonresponse errors verbally but not quantitatively. The news releases and technical documentation published by government statistical agencies caution readers that point estimates of official statistics are subject to nonresponse and other nonsampling errors. However, the estimates themselves contain no information on the potential magnitude of the errors. The documents published by statistical agencies neither specify nor attempt to justify the assumptions of conditionally random nonresponse used to generate imputations and weights.

Statistical agencies could better inform the public if they were to measure potential errors and report them. This paper has shown how to form interval estimates that face up to nonresponse. Section 2 reviewed maximally credible estimates that make no assumptions about the nature of nonresponse. The illustrative applications of Section 3 made plain that nonresponse creates a potentially severe problem for interpretation of American official statistics on income and unemployment.

The most appealing ways to mitigate the identification problem created by nonresponse are to improve the response rates of our surveys and to obtain validation data that shed light on the nature of nonresponse. Short of this, the only way to tighten the interval estimates shown in Section 3 is to make assumptions that either directly or indirectly constrain the distribution of missing data. Section 4 posed various assumptions that may yield narrower intervals. I did not opine on whether any of these or other assumptions is sufficiently credible that agencies should feel comfortable using them when reporting particular statistics. I recommend that analysts at statistical agencies should consider carefully the types of assumptions they deem credible enough to maintain, determine their identifying power, and report interval estimates accordingly.

I think it likely that agencies will want to develop separate strategies for communication of uncertainty in their news releases and in their technical documentation of official statistics. News releases

are brief and are aimed at a broad audience. Hence, they have only a limited ability to convey nuance. It will be challenging for agencies to write releases that effectively quantify sampling and nonsampling error, but it is important that they do so.

Agencies have much more scope for communication of uncertainty in their technical documentation of official statistics. Given adequate budget for analysis and dissemination, agencies can quantify error under a spectrum of alternative assumptions and can perform sensitivity analyses as described in Section 4. The possibilities grow as the norm in documentation moves away from publication of periodic reports and towards development of web-based software that enables the users of official statistics to flexibly access and study the available data.

References

- Horowitz, J. and C. Manski (1995), "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63, 281–302.
- Horowitz, J. and C. Manski (1998), "Censoring of Outcomes and Regressors due to Survey Nonresponse: Identification and Estimation Using Weights and Imputations," *Journal of Econometrics*, 84, 37–58.
- Horowitz, J. and C. Manski (2000), "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95, 77–84.
- Imbens, G. and C. Manski (2004), "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857.
- Kline, P. and A. Santos (2013), "Sensitivity to Missing Data Assumptions: Theory and an Evaluation of the U.S. Wage Structure," *Quantitative Economics*, 4, 231–267.
- Manski, C. (1989), "Anatomy of the Selection Problem," *Journal of Human Resources*, 24, 343–360.
- Manski, C. (1994), "The Selection Problem. "in C. Sims ed. *Advances in Econometrics, Sixth World Congress*, Cambridge: Cambridge University Press.
- Manski, C. (2003), *Partial Identification of Probability Distributions*, New York: Springer-Verlag.
- Manski, C. (2007), *Identification for Prediction and Decision*, Cambridge: Harvard University Press.
- Manski, C. (2013), *Public Policy in an Uncertain World*, Cambridge, MA: Harvard University Press.
- Robins, J., A. Rotnitzky, and D. Scharfstein (1999), "Sensitivity Analysis for Selection Bias and Unmeasured Confounding in Missing Data and Causal Inference Models," in *Statistical Models in Epidemiology*, E. Halloran and D. Berry (editors), New York: Springer, 1–94.
- Rosenbaum, P. (2002), *Observational Studies*, New York: Springer.
- Stoye, J. (2010), "Partial Identification of Spread Parameters," *Quantitative Economics*, 1, 323–357.
- U. S. Bureau of Labor Statistics (2012), *Employment Situation News Release*, October 5, www.bls.gov/news.release/archives/empst 10052012.htm, accessed November 28, 2012.
- U. S. Census Bureau (2006), *Current Population Survey Design and Methodology*, Technical Paper 66, Washington, DC: U. S. Census Bureau.
- U.S. Census Bureau (2011), *Current Housing Reports, Series H150/09, American Housing Survey for the United States: 2009*, Washington, DC: U.S. Government Printing Office.
- U. S. Census Bureau (2012A), *Income, Poverty and Health Insurance Coverage in the United States: 2011*, September 12, www.census.gov/newsroom/releases/archives/income_wealth/cb12-172.html, accessed

November 28, 2012.

U. S. Census Bureau (2012B), *Income, Poverty, and Health Insurance Coverage in the United States: 2011*, Publication P60_243, <http://www.census.gov/prod/2012pubs/p60-243.pdf>, accessed June 12, 2014.

U. S. Census Bureau (2012C), *Source and Accuracy of Estimates for Income, Poverty, and Health Insurance Coverage in the United States: 2011*, Publication P60_243sa, www.census.gov/hhes/www/p60_243sa.pdf, accessed November 28, 2012.

TABLE 1: Interval Estimates Recognizing Item Nonresponse

A. Annual Median Household Income				
Year	Number of Interviewed Households	Fraction with Missing Data	Interval Estimate	Point Estimate with Imputations
2001	78265	0.461	[32000, 102000]	44200
2002	78310	0.466	[31225, 104612]	44231
2003	77149	0.464	[32040, 106000]	45100
2004	76447	0.463	[33000, 106847]	46324
2005	75939	0.440	[35600, 101000]	48078
2006	75477	0.423	[36462, 99999]	50002
2007	75872	0.428	[38000, 101747]	52000
2008	76185	0.417	[39157, 100100]	52004
2009	76260	0.416	[37481, 99999]	51157
2010	75188	0.414	[36909, 100000]	51016
2011	74838	0.408	[38000, 100000]	52000

B. Annual Family Poverty Rate				
Year	Number of Interviewed Families	Fraction with Missing Data	Interval Estimate	Point Estimate with Imputations
2001	89063	0.436	[0.110, 0.315]	0.146
2002	89098	0.440	[0.112, 0.328]	0.152
2003	87948	0.438	[0.117, 0.331]	0.158
2004	87149	0.437	[0.118, 0.331]	0.160
2005	86882	0.415	[0.121, 0.319]	0.161
2006	86222	0.400	[0.120, 0.323]	0.154
2007	86955	0.403	[0.120, 0.324]	0.154
2008	87562	0.392	[0.126, 0.320]	0.162
2009	88957	0.388	[0.138, 0.338]	0.176
2010	87076	0.390	[0.138, 0.344]	0.178
2011	86038	0.384	[0.139, 0.339]	0.176

C. March Unemployment Rate				
Year	Number of Interviewed Civilian Adults	Fraction with Missing Data	Interval Estimate	Point Estimate with Imputations
2002	108901	0.0028	[0.057, 0.062]	0.058
2003	110375	0.0032	[0.060, 0.065]	0.061
2004	108221	0.0030	[0.057, 0.062]	0.057
2005	106913	0.0035	[0.052, 0.057]	0.052
2006	106234	0.0030	[0.047, 0.051]	0.047
2007	105392	0.0039	[0.044, 0.050]	0.044
2008	105643	0.0047	[0.049, 0.057]	0.050
2009	106923	0.0052	[0.085, 0.093]	0.086
2010	107582	0.0053	[0.095, 0.103]	0.096
2011	105774	0.0072	[0.085, 0.096]	0.086
2012	105314	0.0076	[0.078, 0.090]	0.079

TABLE 2: Interval Estimates Recognizing Item and Unit Nonresponse

A. Annual Median Household Income				
Year	Number of Households in Sample	Fraction with Unit Nonresponse	Fraction with Item Nonresponse	Interval Estimate
2001	84831	0.077	0.425	[28200, 123643]
2002	85092	0.080	0.429	[27141, 104611]
2003	84116	0.083	0.425	[28000, 106000]
2004	83932	0.089	0.422	[28240, 106487]
2005	83009	0.085	0.403	[30652, 101000]
2006	82554	0.086	0.387	[31309, 99999]
2007	82235	0.077	0.395	[33200, 95008]
2008	81904	0.070	0.388	[35000, 100100]
2009	81938	0.069	0.387	[33004, 99999]
2010	81737	0.080	0.381	[31800, 100000]
2011	81573	0.088	0.372	[32132, 100000]

B. Annual Family Poverty Rate				
Year	Number of Families in Sample	Fraction with Unit Nonresponse	Fraction with Item Nonresponse	Interval Estimate
2001	95629	0.069	0.406	[0.102, 0.362]
2002	95880	0.071	0.409	[0.104, 0.376]
2003	94915	0.073	0.406	[0.108, 0.380]
2004	94634	0.079	0.403	[0.109, 0.384]
2005	93952	0.075	0.384	[0.112, 0.370]
2006	93299	0.076	0.370	[0.111, 0.374]
2007	93318	0.068	0.376	[0.112, 0.370]
2008	93281	0.061	0.368	[0.118, 0.362]
2009	94635	0.060	0.365	[0.130, 0.378]
2010	93625	0.070	0.363	[0.128, 0.390]
2011	93228	0.077	0.354	[0.128, 0.390]

C. March Unemployment Rate				
Year	Number of Civilian Adults in Sample	Fraction with Unit Nonresponse	Fraction with Item Nonresponse	Interval Estimate
2002	113963	0.044	0.0027	[0.053, 0.123]
2003	115116	0.041	0.0031	[0.056, 0.123]
2004	113461	0.046	0.0028	[0.054, 0.126]
2005	112648	0.051	0.0033	[0.048, 0.129]
2006	111620	0.048	0.0029	[0.043, 0.120]
2007	111028	0.051	0.0037	[0.040, 0.122]
2008	110761	0.046	0.0045	[0.046, 0.122]
2009	111541	0.041	0.0050	[0.073, 0.149]
2010	112190	0.041	0.0051	[0.089, 0.159]
2011	111061	0.048	0.0069	[0.078, 0.162]
2012	111121	0.052	0.0072	[0.071, 0.162]