12-2006

Partially Identifying the Prevalence of Health Insurance Given Contaminated Sampling Response Error

Brent Kreider
Iowa State University, bkreider@iastate.edu

Recommended Citation

Kreider, Brent, "Partially Identifying the Prevalence of Health Insurance Given Contaminated Sampling Response Error" (2006). *Economics Working Papers (2002–2016)*. 263.
https://lib.dr.iastate.edu/econ_las_workingpapers/263

Partially Identifying the Prevalence of Health Insurance Given Contaminated Sampling Response Error

Abstract

This paper derives simple closed-form identification regions for the U.S. nonelderly population's prevalence of health insurance coverage in the presence of household reporting errors. The methods extend Horowitz and Manski's (1995) nonparametric analysis of contaminated samples for the case that the outcome is binary. In this case, draws from the alternative distribution (i.e., not the distribution of interest) might naturally be defined as response errors. The derived identification regions can dramatically reduce the degree of uncertainty about the outcome distribution compared with the contaminated sampling bounds. These regions are estimated using data from the Medical Expenditure Panel Survey (MEPS) combined with health insurance validation data available for a nonrandom portion of the sample.

Keywords

contaminated sampling, partial identification, nonparametric bounds, classification error

Disciplines

Econometrics | Health Economics | Health Policy

IOWA STATE UNIVERSITY

Partially Identifying the Prevalence of Health Insurance Given Contaminated Sampling Response Error

Brent Kreider

April 2006

Working Paper # 06017

Department of Economics
Working Papers Series

Ames, Iowa 50011

Partially Identifying the Prevalence of Health Insurance Given Contaminated Sampling Response Error

Brent Kreider*
Department of Economics
Iowa State University
bkreider@iastate.edu

December 2006

Abstract. This paper derives simple closed-form identification regions for the U.S. nonelderly population's prevalence of health insurance coverage in the presence of household reporting errors. The methods extend Horowitz and Manski's (1995) nonparametric analysis of contaminated samples for the case that the outcome is binary. In this case, draws from the alternative distribution (i.e., not the distribution of interest) might naturally be defined as response errors. The derived identification regions can dramatically reduce the degree of uncertainty about the outcome distribution compared with the contaminated sampling bounds. These regions are estimated using data from the Medical Expenditure Panel Survey (MEPS) combined with health insurance validation data available for a nonrandom portion of the sample.

JEL classification numbers: C14, C21, I18

Keywords: contaminated sampling, partial identification, nonparametric bounds, classification error

*The author gratefully acknowledges financial support by a grant from the Robert Wood Johnson Foundation through the Economic Research Initiative on the Uninsured (ERIU) PO#3000370614.

1 Introduction

Economic analyses of the nation's access to health care and the cost of covering the uninsured inevitably require estimates of the population's prevalence of health insurance coverage (Institute of Medicine 2003). Rhoades (2005) reports that about a quarter of the U.S. nonelderly population was uninsured during at least part of 2003. Recent evidence, however, suggests that health insurance data collected in household surveys may be prone to substantial reporting error. Using matched surveys of employers and their employees, for example, Berger, Black, and Scott (2000) report that more than one-fifth of the workers and their employers disagreed about whether the worker was eligible for insurance. Nelson et al. (2000) find large inconsistencies in reported source and duration of coverage.[1]

Survey respondents may be unaware of their own current insurance status or that of a family member, and they may imperfectly recall past coverage. Several researchers describe widespread potential for insurance classification errors (e.g., Swartz, 1986; Czajka and Lewis, 1999; Monheit, 2003; Short, 2004). Observed inconsistencies have become a "source of confusion" to researchers and policymakers seeking to obtain reliable indicators of the size of the uninsured population (Monheit, 2003). The Census Bureau now cautions against drawing strong inferences from insurance data collected by the Current Population Survey (CPS), the official source of health insurance statistics in the United States (DeNaves-Walt *et al.*, 2005).

Using a nonparametric partial identification framework, this paper is the first to investigate what can be identified about the nonelderly population's prevalence of health insurance coverage when the data may be contaminated with classification errors. In Horowitz and Manski's (H-M, 1995) seminal work, an outcome is drawn from a mixture of a distribution of interest, F, and an alternative distribution, G. The researcher has no information about whether a particular draw comes from F or G. Draws from F are known to be accurate, while draws from G may be either accurate or inaccurate.[2] H-M consider, for example, the case that contamination of income values in microdata

[1] For example, when an insurer said a respondent was insured for a year or less, the respondent's report agreed only 40% of the time.

[2] Although H-M refer to realizations from the alternative distribution as constituting data errors, they do not require that all realizations from that unknown distribution are erroneous.

arises from imputation. Imputed values are treated as unreliable, and the researcher does not know which values are imputed. The most conservative "corrupt sampling" environment imposes only a lower bound on the fraction of draws coming from F. The "contaminated sampling" environment imposes an additional assumption that observations are drawn from F or G independently of the realized value of the draw. H-M provide sharp bounds on the outcome distribution of interest in both of these environments.

This paper extends the H-M analysis for the case that the outcome variable – in this case health insurance status – is binary. When the outcome is binary, draws from the alternative distribution might naturally be defined as response errors. In that case, draws from F are known to be accurate and draws from G are known to be erroneous. Sharp identification regions for the prevalence of health insurance are derived in this random errors environment and compared with the H-M contaminated sampling bounds. The sensitivity of the results to departures from strict independence is also examined. The analytical results contribute to the recent literature on classification error in binary variables (e.g., Bollinger, 1996; Berger, Black, and Scott, 2000; Frazis and Loewenstein, 2003; Bollinger and David, 1997, 2001, 2005) and to the literature on measurement error in corrupt and contaminated samples (e.g., Horowitz and Manski, 1995; Lambert and Tierney, 1997; Pepper, 2000; Dominitz and Sherman, 2004, 2006; Molinari, 2005; and Kreider and Pepper, forthcoming).

2 Data

The data come from the 1996 Medical Expenditure Panel Survey (MEPS), a nationally representative household survey conducted by the Agency for Healthcare Research and Quality (AHRQ). The MEPS goes to great lengths to elicit accurate health insurance information from all potential sources, and these data arguably provide the most accurate information about insurance coverage in the U.S.[3] Since nearly all adults become eligible for Medicare when they turn 65, this paper focuses on the nonelderly population; the sample consists of 18,851 children and adults younger

[3] In a detailed series of questions, respondents were queried about public coverage through Medicare, Medicaid, Champus/Champva, or any other government agency that provided hospital and physician benefits. They were also asked about private coverage from a current or previous employer, insurance company, union, or any other group or association. Where applicable, the survey used state-specific program names to aid in recognition.

than 65 in July 1996.

For the month of July 1996, the MEPS conducted follow-back interviews with employers, unions, and insurance companies that can be used to corroborate self-reported insurance status for a non-random portion of the sample. Respondents were also asked to show insurance cards and policy booklets. Based on reported insurance status by family respondents, 81% of the nonelderly population was insured in July 1996 and 19% was uninsured. Kreider and Hill (2006) use this sample to study health care utilization. As described in that paper, insurance status can be "verified" for 67% of the sample. True insured status for the remaining observations is unconfirmed and subject to classification error. While there is no gold standard for the accuracy of insurance status (employers and insurance companies could also be mistaken, etc.), the verification approach represents a compromise between taking all self-reported data at face value and completely discarding the data.

3 The Identification Problem

Let $I^* = 1$ indicate that a person is truly insured, with $I^* = 0$ otherwise. We observe the self-reported counterpart I. A latent variable Z^* indicates whether a classification comes from the distribution of interest, F, or the alternative distribution, G. In the H-M framework, draws from F are known to be accurate while draws from G may be either accurate or inaccurate. The objective is to learn about the distribution of I^*. Given that I^* is binary in this analysis, we are interested in learning about $P(I^* = 1)$.[4] In the most conservative corrupt sampling environment, the only source of knowledge is a presumed lower bound on the fraction of draws coming from F. Under contaminated sampling, H-M additionally impose the independence assumption $I^* \perp Z^*$.

This paper considers the additional identifying power of distinguishing draws from F and G as being accurate and inaccurate, respectively. In this case, Z^* is defined to equal 1 if I and I^* coincide, with $Z^* = 0$ otherwise. In a binary setting, knowledge that draws from the alternative distribution are inaccurate can have identifying power compared with the H-M environment because the event $Z^* = 0$ perfectly identifies the true value of the outcome: $I^* = 1 - I$. Under corrupt sampling, the redefinition of Z^* has no implications for what can be identified about the outcome

[4] Like H-M, the notation leaves implicit any conditioning variables of interest.

distribution. Under contaminated sampling, however, defining draws from G as response errors has identifying power because, in that case, the assumption $I^* \perp Z^*$ implies that classification errors arise independently of the value of I^*. This paper derives sharp identification regions on $P(I^* = 1)$ in this "orthogonal errors" setting. Except in special cases, these regions lie strictly inside the H-M contaminated sampling bounds (in the special cases, they are identical).

As described in Section 2, some respondents' insurance responses have been corroborated by other sources. Let $Y = 1$ indicate that a response I is verified to be accurate (i.e., Z^* is known to equal 1). If $Y = 0$, then Z^* may be either 1 or 0.[5] Using the law of total probability, the true insured rate can be decomposed as

$$P(I^* = 1) = P(I^* = 1|Y = 1)P(Y = 1) + P(I^* = 1|Y = 0)P(Y = 0). \tag{1}$$

Each of the terms in (1) is identified except for the third term. We observe the true insured rate among verified responses, $P(I^* = 1|Y = 1)$, and we observe the fractions of verified and non-verified responses. The population's insured rate is not identified, however, because we do not observe the true insured rate among unverified cases.

The H-M corrupt sampling environment considers what can be identified when an arbitrary fraction of unverified outcomes may be arbitrarily mismeasured. The only assumption is a limit on the degree of potential data corruption:

$$P(Z^* = 1|Y = 0) \geq v. \tag{2}$$

If $v = 1$, as implicitly assumed in most studies, then all insurance classifications are known to be accurate: $P(I^* = 1|Y = 0) = P(I = 1|Y = 0)$. Following H-M and the literature on robust statistics (e.g., Huber, 1981), we can study the path of identification decay as v departs from 1; "identification breakdown" occurs at the largest value of v such that we can no longer obtain informative lower and upper bounds on $P(I^* = 1|Y = 0)$.

Most of the econometric literature assessing classification error presumes that the majority of potentially misclassified responses are accurate (implying $v > \frac{1}{2}$). That is, the classifications in the

[5] H-M implicitly assume that $Y = 0$ for all observations. In analyses of testing for environmental pollutants and evaluating school performance, Dominitz and Sherman (2004, 2006) were the first to distinguish between "verified" and "unverified" observations in the data.

data are presumed to be more informative than their converse (e.g., Bollinger, 1996). Using internal MEPS data at the Agency for Healthcare Research and Quality, Hill (2006) conducts a detailed exploration of the accuracy of self-reported insurance status in the MEPS. His analysis proposes two candidate values of v for the July 1996 sample, 0.74 and 0.95, depending on the maintained assumptions.[6] If a researcher has no confidence in the self-reported data, then v can be set equal to 0. Under corruption or contamination in the H-M framework, we can learn nothing about the outcome distribution for sufficiently small values of v. In the Proposition 1 bounds below, however, we can place informative bounds on $P(I^* = 1|Y = 0)$ even when $v = 0$ except in the special case that $P(I = 1|Y = 0) = \frac{1}{2}$.

The next section investigates how assumptions on the reporting error process in a binary setting translate into restrictions on patterns of false positive and false negative classifications. These restrictions lead to sharp identification regions for the true insured rate.

3.1 Derivation of sharp identification regions

Denote $\theta^+ \equiv P(I = 1, Z^* = 0|Y = 0)$ and $\theta^- \equiv P(I = 0, Z^* = 0|Y = 0)$ the unknown proportions of false positives and false negatives among unverified cases, respectively, and denote $p \equiv P(I = 1|Y = 0)$ the reported insured rate. The objective is to deduce identification regions for the true insured rate among unverified cases,

$$P^* \equiv P(I^* = 1|Y = 0) = p + \theta^- - \theta^+, \tag{3}$$

the unidentified component in (1).

Equation (2) implies the following restrictions on θ^+ and θ^-:

$$(i) \qquad 0 \leq \theta^+ \leq \min\{1 - v,\ p\} \tag{4}$$

$$(ii) \qquad 0 \leq \theta^- \leq \min\{1 - v, 1 - p\} \tag{5}$$

$$(iii) \qquad 0 \leq \theta^+ + \theta^- \leq 1 - v. \tag{6}$$

That is, the fraction of false positive responses cannot exceed the total fraction of positive responses, p, or the maximum allowed fraction of total misclassifications, $1 - v$. Similarly, the fraction of false

[6] The larger value relies on strong assumptions about the degree to which validation information for a subset of the observations can be extrapolated to the remainder of the sample. The smaller value relies on weaker assumptions.

negative responses cannot exceed the total fraction of negative responses, $1-p$, or the maximum allowed fraction of total misclassifications. Finally, the sum of the false positive and false negative classifications cannot exceed the maximum allowed fraction of total misclassifications.

We can also consider the power of the "orthogonal errors" restriction described above:

$$P(I^*=1|Z^*=0,Y=0)=P(I^*=1|Z^*=1,Y=0). \tag{7}$$

This independence assumption implies a quadratic relationship between θ^+ and θ^-:[7]

$$(iv) \qquad \left(\theta^-\right)^2-(1-p)\theta^-+\theta^+\left(p-\theta^+\right)=0. \tag{8}$$

These constraints are illustrated in Figure 1. In the MEPS sample, $p=0.485$. Constraints (i) and (ii) restrict combinations of $\left\{\theta^+,\theta^-\right\}$ to lie within the rectangle $oef'a$ with width $p=0.485$ and height $1-p=0.515$. Along the diagonal oo', θ^+ and θ^- exactly cancel out such that the true insured rate equals the reported rate: $P^*=p$. The true insured rate is also constant along any diagonal parallel to oo' and falls as we consider parallels further to the right. For illustration, v is set to 0.7 in the figure which implies that $\left\{\theta^+,\theta^-\right\}$ must also lie within the triangle occ'. This triangle shrinks with v as the degree of confidence in the data rises, and it expands as the degree of confidence declines.

Under corrupt sampling, constraint (iv) does not apply. In this case, we obtain the H-M (Corollary 1.2) corrupt sampling lower bound by setting $\theta^-=0$ and $\theta^+=\min\{1-v,\ p\}$ and obtain the upper bound by setting $\theta^+=0$ and $\theta^-=\min\{1-v,\ 1-p\}$:

$$LB^{HM}_{corrupt}=\max\{0,\ p-(1-v)\}\leq P^*\leq\min\{1,\ p+(1-v)\}=UB^{HM}_{corrupt}. \tag{9}$$

Applying the H-M contaminated sampling bounds (H-M Corollary 1.2) to a binary outcome obtains the tighter bounds

$$LB^{HM}_{contam}\equiv\max\left\{0,\frac{p-(1-v)}{v}\right\}\leq P^*\leq\min\left\{1,\frac{p}{v}\right\}\equiv UB^{HM}_{contam}. \tag{10}$$

These bounds can be tightened further under the orthogonal errors assumption in (7). In particular, constraint (iv) rules out all $\left\{\theta^+,\theta^-\right\}$ that do not satisfy

$$\theta^-(\theta^+)=\frac{(1-p)\pm\sqrt{4(\theta^+)^2-4p\theta^++(1-p)^2}}{2}. \tag{11}$$

[7] To see this for the case $\theta^++\theta^-\in(0,1)$, write the independence restriction (7) as $\frac{\theta^-}{\theta^++\theta^-}=\frac{p-\theta^+}{1-\theta^+-\theta^-}$.

The curved lines in Figure 1 trace out combinations of $\{\theta^+, \theta^-\}$ that satisfy (11) for the case $p < \frac{1}{2}$. Curves oa and ef' represent the smaller and larger roots of θ^-, respectively.[8] Combinations of $\{\theta^+, \theta^-\}$ outside the rectangle $oef'a$ involve imaginary roots and are already ruled out by constraints (*i*) and (*ii*). It is straightforward to show that curve oa is concave with slope everywhere less than one, and curve ef' is convex with slope everywhere greater than -1. Starting at point a with $\{\theta^+ = p,\, \theta^- = 0\}$, the true insured rate $P^* = p + \theta^- - \theta^+$ equals 0. Moving to the left along the oa curve, P^* rises continuously to the reported rate p at point o. For $P^* \in (p, 1-p)$ (inside the region $off'o'$), no combinations of $\{\theta^+, \theta^-\}$ satisfy independence. Values of $P^* \in [1-p, 1]$ consistent with independence lie on curve ef'. When all classifications are inaccurate at point f' (i.e., $\theta^+ = p$ and $\theta^- = 1-p$), we have $P^* = 1-p$. Moving to the left along the curve, everyone is insured at point e ($\theta^+ = 0$ and $\theta^- = 1-p$).

We can now assess identification across values of v. If $v = 1$, then $P^* = p$ at point o. Next consider $v \in [1-p, 1)$. Since P^* strictly declines as we move to the right along curve oa from point o ($P^* = p$) to point a ($P^* = 0$), the upper bound on P^* is given by p. The lower bound is determined by $p + \theta_v^- - \theta_v^+$ at point b (illustrated for $v = 0.7$), where θ_v^+ and θ_v^- must simultaneously satisfy $\theta^- + \theta^+ = 1 - v$ and (11). The unique solution is $P^* = \frac{v-(1-p)}{2v-1}$. For $v \in (p, 1-p]$, P^* must lie within $[0,\, p]$. For $v \in [0, p]$, P^* must lie within $[0,\, p] \cup \left[\frac{v-(1-p)}{2v-1}, 1\right]$. When $p = \frac{1}{2}$, nothing is known about P^* unless $v > \frac{1}{2}$, in which case $P^* = p$. After deriving analogous results for $p > \frac{1}{2}$, we obtain the following identification regions for P^*:

Proposition 1. *Suppose $P(Z^* = 1|Y = 0) \geq v$ and the orthogonal errors assumption holds. Then $P^* \equiv P(I^* = 1|Y = 0)$ lies within the following regions:*

$$P^* \in \begin{cases} [0,\, p] \cup \left[\frac{v-(1-p)}{2v-1}, 1\right] & \text{if } p < \frac{1}{2} \text{ and } v \in [0,\, p] \\ [0,\, p] & \text{if } p < \frac{1}{2} \text{ and } v \in (p,\, 1-p] \\ \left[\frac{v-(1-p)}{2v-1},\, p\right] & \text{if } p < \frac{1}{2} \text{ and } v \in (1-p,\, 1] \end{cases} \tag{12}$$

$$P^* \in \begin{cases} [0,\, \frac{v-(1-p)}{2v-1}] \cup [p, 1] & \text{if } p > \frac{1}{2} \text{ and } v \in [0,\, 1-p] \\ [p,\, 1] & \text{if } p > \frac{1}{2} \text{ and } v \in (1-p,\, p] \\ \left[p,\, \frac{v-(1-p)}{2v-1}\right] & \text{if } p > \frac{1}{2} \text{ and } v \in (p,\, 1] \end{cases} \tag{13}$$

$$P^* \in \begin{cases} [0,\, 1] & \text{if } p = \frac{1}{2} \text{ and } v \in \left[0, \frac{1}{2}\right] \\ \{p\} & \text{if } p = \frac{1}{2} \text{ and } v \in (\frac{1}{2}, 1] \end{cases} \tag{14}$$

[8] As p gets smaller, the curves flatten toward the horizontal lines oa and ef'.

A proof is provided in the appendix.

There are several notable features of these regions. First, it is straightforward to check that these regions lie strictly inside the H-M contaminated sampling bounds in (10) for $v < 1$ unless $v \leq p = \frac{1}{2}$ (in which case both sets of bounds are uninformative: $P^* \in [0, 1]$). For example, suppose $v > p > \frac{1}{2}$. Under the H-M corrupt sampling bounds in (9), P^* can be as large as $p + (1 - v)$. The upper bound declines to $\frac{p}{v}$ using the H-M contaminated sampling bounds in (10) and declines further to $\frac{v-(1-p)}{2v-1}$ using Proposition 1. Moreover, the region is not contiguous for $v \leq \min\{p, 1-p\}$. As discussed earlier, however, much of the literature assessing classification error presumes that the majority of the classifications are accurate ($v > \frac{1}{2}$). That assumption implies $v > \min\{p, 1-p\}$, thus ruling out disjoint regions for that case. Note also that for a sufficiently large v, P^* is bounded above (below) by p if $p < \frac{1}{2}$ ($p > \frac{1}{2}$).

Introducing a linear programming approach affordable to a general class of models, Molinari (2005, Proposition 7) has independently derived comparable regions for the case of "constant probability of correct report." Her highly technical "direct misclassification" approach exploits the fact that relationships between the distribution of a true variable and its mismeasured counterpart can be represented by a linear system of simultaneous equations involving a coefficient matrix of misclassification probabilities. She shows how restrictions on this matrix can be used to derive identification regions for unknown parameters of interest. The current analysis derives identification regions in the H-M contaminated sampling setting, making explicit and transparent how a response error identifying assumption translates into restrictions on false positive and false negative classifications $\{\theta^+, \theta^-\}$. It also investigates how the set of feasible combinations of $\{\theta^+, \theta^-\}$ expands as the independence assumption is relaxed; in the limit as independence becomes fully relaxed, the Proposition 1 identification regions expand to the H-M corrupt sampling bounds in (9).

4 Empirical results

Table 1 and Figure 2 present bounds on the nonelderly population's true insured rate, $P^o \equiv P(I^* = 1)$. When $v = 1$, P^o is point-identified as the nonelderly population's self-reported insured rate, 0.807. Allowing for reporting errors, we can examine the rate of identification decay as v departs from 1. When nothing is known about the patterns of errors (corrupt sampling), each percentage point decline in v results in a 6.6 percentage point increase in the width of the estimated bounds on P^o.[9]

Under Hill's (2006) conservative proposed value $v = 0.74$, the true insured rate lies within $[0.722, 0.893]$ under corrupt sampling after accounting for sampling variability, a 17 percentage point range of uncertainty.[10] The H-M contaminated sampling bounds reduce this uncertainty by 5.5 percentage points to $[0.748, 0.864]$, a 5 point reduction in the width. In contrast, the Proposition 1 response error bounds nearly point-identify the true insured rate: $P^o \in [0.802, 0.807]$, with width less than 1/20th that of the H-M contaminated sampling bounds. When $v = 0.95$, the Proposition 1 bounds are about half as wide as the H-M contaminated sampling bounds.

For v less than $p = 0.485$, the Proposition 1 identification region is disjoint. When $v = 0.4$, for example, the true insured rate lies within the region $[0.648, 0.807] \cup [0.838, 0.977]$. This region is equivalent to the H-M contaminated sampling bounds $[0.648, 0.977]$ except that interior values between 0.807 and 0.838 are infeasible.[11] In contrast to the H-M bounds, the Proposition 1 regions have identifying power for all values of v.

We can examine the sensitivity of these results to departures from the reference case of strict independence. The orthogonality assumption (7) can be written as $P(I^* = 1|Z^* = 0, Y = 0) = \kappa P(I^* = 1|Z^* = 1, Y = 0)$ with $\kappa = 1$. Among unverified cases, the true insured rate is identical among accurate and inaccurate reporters. Relaxing this assumption, suppose instead that κ is allowed to lie anywhere within the range $[\kappa_1, \kappa_2]$. Figure 3 illustrates how the identification regions expand under "partial independence." When $\kappa \in [0.9, 1.1]$ in Case (a), the true insured rate among inaccurate reporters is allowed to deviate up to 10% from that among accurate reporters. In this case, the bounds are identical to the H-M corrupt sampling bounds (and contaminated sampling bounds) for low values of v and depicted by the Case (a) dotted lines for higher values of v. There are no disjoint regions in this case; the bounds resemble the H-M contaminated sampling bounds

[9] For sufficiently large values of v, the slope of the corrupt sampling lower and upper bound is 0.33 and -0.33, respectively, resulting from $P(Y = 0) = 0.33$. Once v falls below $1 - p = 0.515$, the lower bound on P^o cannot fall any further because the lower bound on $P(I^* = 1|Y = 0)$ attains 0; once v falls below $p = 0.485$, the upper bound cannot rise any further.

[10] With nearly 20,000 observations, the uncertainty associated with sampling variability is very small compared with uncertainty associated with the identification problem. Fifth percentile lower bounds and ninety-fifth percentile upper bounds for the identification regions were computed using balanced replicate methods (Wolter, 1985).

[11] As noted above, Proposition 1 rules out disjoint identification regions under the common assumption that the majority of unverified classifications are accurate.

in Figure 2, except they are generally tighter. These bounds widen in Cases (b) and (c) as the independence assumption is further relaxed. In the limit as independence is relaxed completely, the bounds attain the H-M corrupt sampling bounds.

For Case (a), the expansion in the width of the bounds is nearly negligible at $v = 0.95$. Even when $v = 0.74$, the Proposition 1 bounds widen only 2 percentage points, from $[0.802, 0.807]$ to $[0.790, 0.815]$, compared with strict independence. In fact, the response error bounds are sufficiently informative that they remain strictly inside the H-M contaminated sampling bounds, $[0.748, 0.864]$, as long as the true insured rate among inaccurate reporters is anywhere between 43% and 139% of the rate among accurate reporters. That is, there is room to dramatically weaken the orthogonality assumption before the sharp bounds become less informative than the H-M contaminated sampling bounds.

5 Conclusion

In recent years, many researchers have called into question the reliability of household responses to questions about insurance status. Highlighting surprising degrees of insurance classification error in many popular national surveys along with dramatic inconsistencies in responses when experimental follow-up insurance questions have been posed, Czajka and Lewis (1999) write:

> "Until we can make progress in separating the measurement error from the reality of uninsurance, our policy solutions will continue to be inefficient, and our ability to measure our successes will continue to be limited."

Using a partial identification framework, this paper investigated what can be identified about the prevalence of health insurance coverage when the data may be contaminated with household reporting errors. Specifically, the analysis derived closed-form identification regions for the reference case of random classification errors and examined the sensitivity of inferences to different assumptions. For a binary outcome, these regions tighten Horowitz and Manski's (1995) contaminated sampling bounds if draws from the alternative distribution are taken to reflect response errors. Compared with the H-M bounds, the identification regions remain more informative even after allowing for substantial departures from independence. The identification regions can be informative in a wide range of interesting topics in the social sciences such as the use of illicit drugs, receipt of welfare benefits, health and disability status, and racial profiling.

In the 1996 MEPS, outside information from insurance cards, policy booklets, and follow-back interviews with employers and insurance companies could corroborate self-reported insurance status for part of the sample. Combining this verification information with the Proposition 1 identification regions, the population's true insured rate can be confined to a small range even given substantial uncertainty about the reliability of unverified reports. If a researcher is not willing to assume anything about the patterns of health insurance reporting errors, then the H-M corrupt sampling bounds apply. Future research into the nature and degree of health insurance reporting error, as called on by the Institute of Medicine (2003), will help researchers make more informed inferences about the population's access to medical services.

References

[1] Berger, M., Black, D., Scott, F. (2000). "Bounding Parameter Estimates With Non-classical Measurement Error." *Journal of the American Statistical Association*, 95: 739–748.

[2] Bollinger, C. (1996). "Bounding Mean Regressions When a Binary Regressor is Mismeasured." *Journal of Econometrics*, 73: 387-99.

[3] Bollinger, C. and M. David. (1997). "Modeling Discrete Choice with Response Error: Food Stamp Participation." *Journal of the American Statistical Association*, 92:827-35.

[4] Bollinger, C. and M. David. (2001). "Estimation With Response Error and Nonresponse: Food Stamp Participation in the SIPP." *Journal of Business and Economic Statistics*, 19: 129-142.

[5] Bollinger, C. and M. David. (2005). "I Didn't Tell, and I Won't Tell: Dynamic Response Error in the SIPP." *Journal of Applied Econometrics*, 20: 563-569.

[6] Czajka, J. and K. Lewis. (1999). "Using Universal Survey Data to Analyze Children's Health Insurance Coverage: An Assessment of Issues." Washington, DC: Mathematica Policy Research, Inc., http://aspe.os.dhhs.gov/health/reports/Survey%20Data.htm.

[7] DeNaves-Walt, C., B.C. Proctor, and C.H. Lee (2005), Income, Poverty, and Health Insurance Coverage in the United States, 2004. U.S. Census Bureau, Current Population Reports P60-229. U.S. Government Printing Office, Washington DC.

[8] Dominitz, J., and R. Sherman (2004), "Sharp Bounds Under Contaminated or Corrupted Sampling With Verification, With an Application to Environmental Pollutant Data," *Journal of Agricultural, Biological and Environmental Statistics*, 9(3), 319-338.

[9] ________ and ________ (2006), "Identification and Estimation of Bounds on School Performance Measures: A Nonparametric Analysis of a Mixture Model with Verification," *Journal of Applied Econometrics*, 21, 1295-1326.

[10] Frazis, H. and M. Loewenstein. (2003). "Estimating Linear Regressions with Mismeasured, Possibly Endogenous, Binary Explanatory Variables," *Journal of Econometrics*, 117, 151-78.

[11] Hill, S. (2006), "The Accuracy of Reported Insurance Status in the MEPS." Working Paper. Agency for Healthcare Research and Quality.

[12] Huber, P. (1981), Robust Statistics, New York: Wiley.

[13] Horowitz, J., and C. Manski (1995), "Identification and Robustness With Contaminated and Corrupted Data," *Econometrica* 63(2), 281-302.

[14] Institute of Medicine. (2003). Hidden Costs, Lost Value: Uninsurance in America. Washington, DC: National Academy Press.

[15] Kreider, B. and J. Pepper (forthcoming), "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors." *Journal of the American Statistical Association.*

[16] Kreider, B., and S. Hill (2006), "Partially Identifying Treatment Effects with an Application to Covering the Uninsured," mimeo, Department of Economics, Iowa State University.

[17] Lambert, D. and L. Tierney. (1997). "Nonparametric Maximum Likelihood Estimation from Samples with Irrelevant Data and Verification Bias," *Journal of the American Statistical Association*, 92: 937-944.

[18] Molinari, F. (2005). "Partial Identification of Probability Distributions with Misclassified Data." Working Paper. Department of Economics, Cornell University.

[19] Monheit, A.C. (2003), "Verifying Lack of Health Insurance in the Medical Expenditure Panel Survey: An Assessment," Report submitted to the Agency for Health Care Research and Quality. New Brunswick, NJ: Rutgers University.

[20] Nelson, D.E., B.L. Thompson, N.J. Davenport, and L.J. Penaloza. (2000). "What People Really Know about Their Health Insurance: A Comparison of Information Obtained from Individuals and Their Health Insurers." *American Journal of Public Health* 90(6): 94-8.

[21] Pepper, J. (2000), "The Intergenerational Transmission of Welfare Receipt: A Nonparametric Bounds Analysis," *Review of Economics and Statistics*, 82(3), 472-288.

[22] Rhoades, J.A. (2005). "The Uninsured in America 1996-2004: Estimates for the Civilian Noninstitutionalized Population Under Age 65." Statistical Brief #84. Rockville, MD: Agency for Healthcare Research and Quality.

[23] Short, P. F. (2004). "Counting and Characterizing the Uninsured." In C. McLaughlin (ed.), Health Policy and the Uninsured. Washington, D.C.: Urban Institute Press.

[24] Swartz, K. (1986). "Interpreting the Estimates from Four National Surveys of the Number of People Without Health Insurance," *Journal of Economic and Social Measurement*, 14, 233-242.

[25] Wolter, K. M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

6 Appendix

Proof of Proposition 1. Using the law of total probability, we can decompose the reported insured rate among unverified cases into reported rates among accurate and inaccurate reporters:

$$p = P(X = 1|Y = 0, Z^* = 1)z^* + P(X = 1|Y = 0, Z^* = 0)(1 - z^*)$$

where $z^* \equiv P(Z^* = 1|Y = 0)$. Using $P(X = 1|Y = 0, Z^* = 1) = P(I^* = 1|Y = 0, Z^* = 1)$ and the orthogonality assumption in (7), it follows that

$$\begin{aligned} p &= P(I^* = 1|Y = 0, Z = 1)\,(2z^* - 1) + (1 - z^*) \qquad (15) \\ &= (2z^* - 1)\,P^* + (1 - z^*). \end{aligned}$$

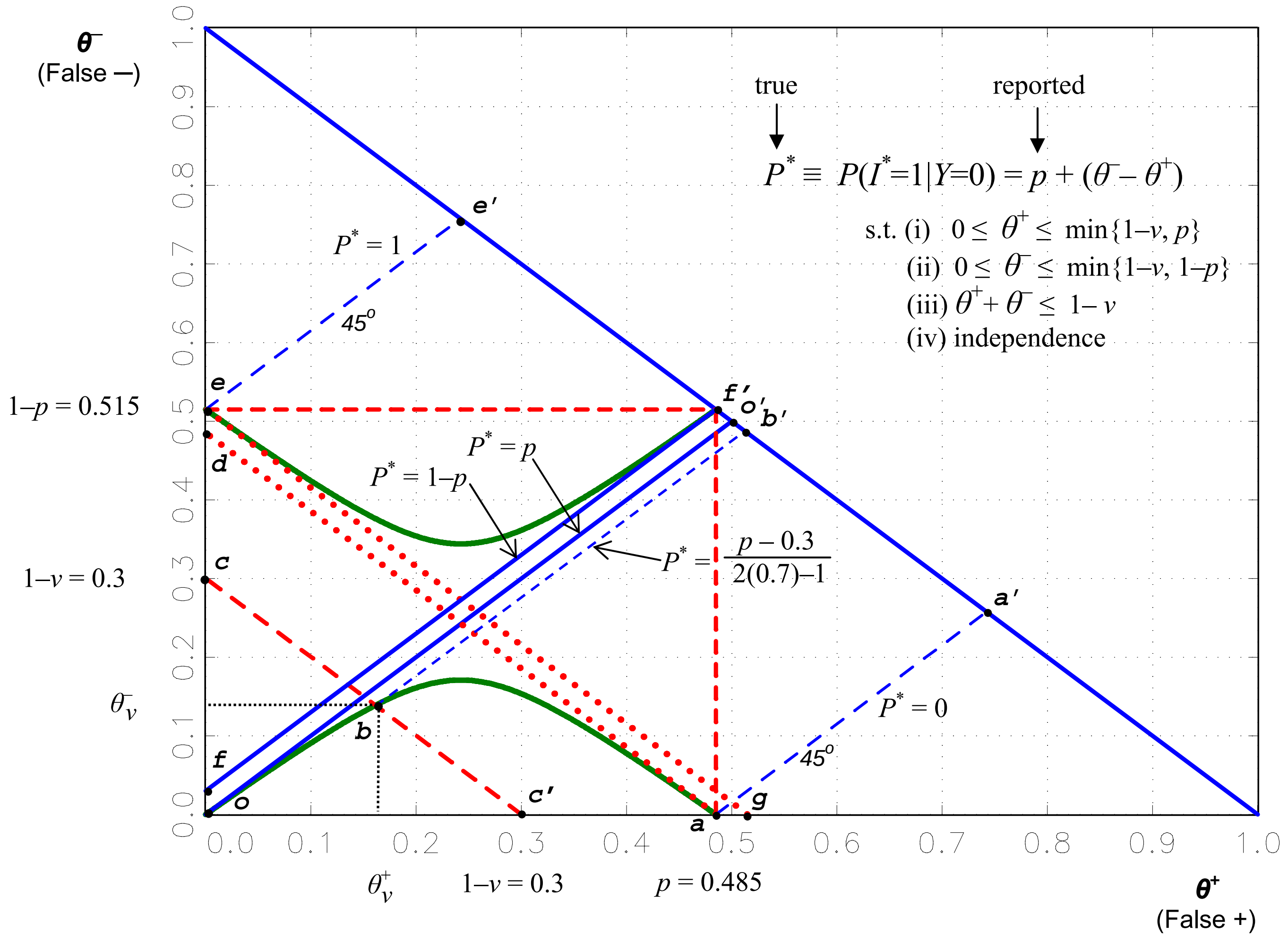
When $z^* = \frac{1}{2}$, p must also equal $\frac{1}{2}$ and nothing is known about P^*. Otherwise,

$$P^*(z^*) = \frac{z^* - (1 - p)}{2z^* - 1} \text{ for } z^* \neq \frac{1}{2}. \qquad (16)$$

When $p < \frac{1}{2}$, $P^*(z^*)$ is strictly increasing from $1-p$ to 1 in the range $[0, p]$, strictly increasing from 0 to p in the range $[1 - p, 1]$, and outside $[0, 1]$ in the range $(p, 1 - p)$. Candidate values of z^* are confined to the space $z^* \in [v, 1] \cap \{[0, p] \cup [1 - p, 1]\}$. When $p > \frac{1}{2}$, $P^*(z^*)$ is strictly decreasing from $1-p$ to 0 in the range $[0, 1 - p]$, strictly decreasing from 1 to p in the range $[p, 1]$, and outside $[0, 1]$ in the range $(1 - p, p)$. Candidate values of z^* are confined to the space $z^* \in [v, 1] \cap \{[0, 1 - p] \cup [p, 1]\}$. When $p = \frac{1}{2}$ and $z^* \neq \frac{1}{2}$, (16) reveals that $P^* = \frac{1}{2}$; when $p = \frac{1}{2}$ and $z^* = \frac{1}{2}$, (15) reveals that P^* can take on any value within $[0, 1]$. These restrictions on P^* establish the identification regions in Proposition 1. □

Figure 1

Constraints on False Positives and False Negatives when Classification
Errors Arise Independently of True Insurance Status



The self-reported insured rate within unverified cases is $p = P(I{=}1|Y{=}0) = 0.485$. The unobserved fraction of false positives on the horizontal axis, $\theta^+ = P(I = 1, Z^* = 0|Y{=}0)$, cannot exceed the total fraction of positive classifications, $p = 0.485$. Similarly, the total fraction of false negatives on the vertical axis, $\theta^- = P(I = 0, Z^* = 0|Y{=}0)$, cannot exceed the total fraction of negative classifications, $1\text{-}p = 0.524$. Therefore, combinations of θ^+ and θ^- are confined to lie within the rectangle *oef′a*. The diagonal *oo′* represents cases in which false positives and false negatives exactly cancel out such that the true insured rate equals the reported rate: $P^* = p$. The true insured rate is constant along any diagonal line parallel to *oo′*. The value of P^* falls as we consider diagonals further to the right.

The curved lines trace out combinations of θ^+ and θ^- that satisfy the independence constraint. Suppose the lower bound accurate reporting rate is $v = 0.7$ (as indicated in the figure). Then since θ^+ and θ^- must lie within the triangle *occ*', they are restricted to lie on the arc *ob*. In this case, the upper bound on P^* is attained at point *o* such that $P^* = p$. The lower bound is attained at point *b* such that $P^* = (p\text{-}0.3)/[2(0.7)\text{-}1]$. As $1\text{-}v$ rises, point *b* moves to the right along the curve and the lower bound continuously declines. For values of $1\text{-}v$ exceeding $1\text{-}p = 0.524$, some combinations of θ^+ and θ^- on the upper curve *ef′* become feasible. Note that combinations of θ^+ and θ^- lying within the region *adeg* are never possible, resulting in the disjoint regions indicated in Proposition 1 and Table 1 for sufficiently small values of v. Disjoint regions are ruled out if the majority of classifications are assumed to be accurate.

Figure 2. Identification Regions for the U.S. Nonelderly Population's True Insured Rate, P^o

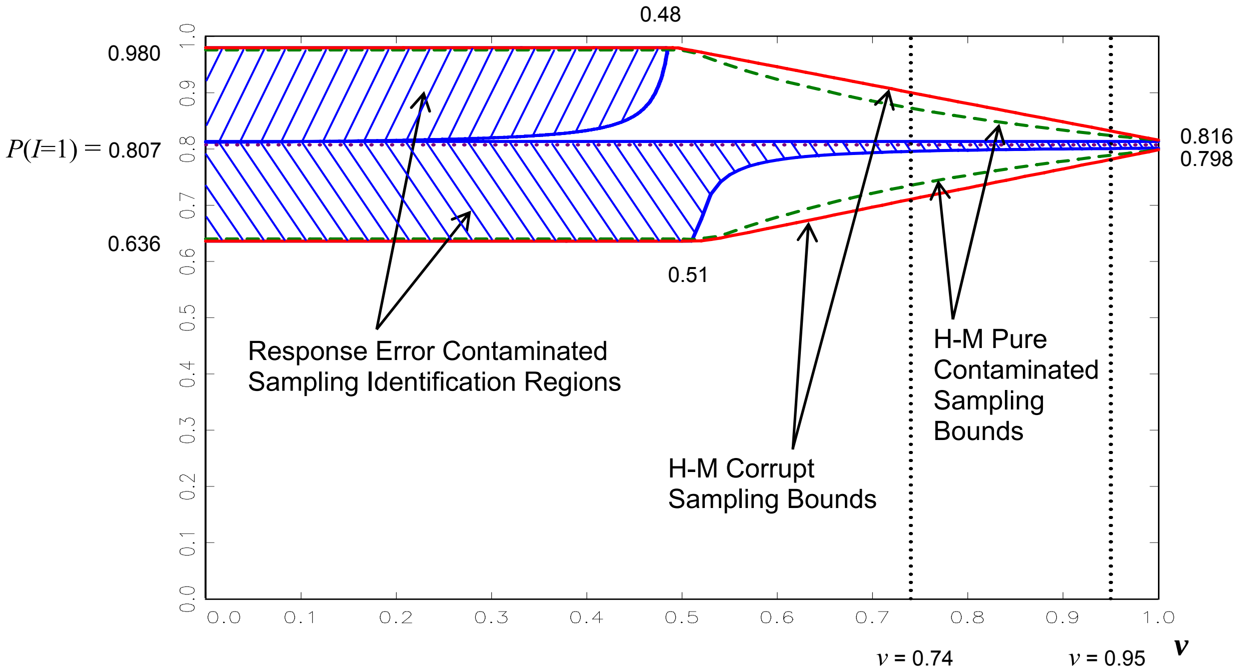
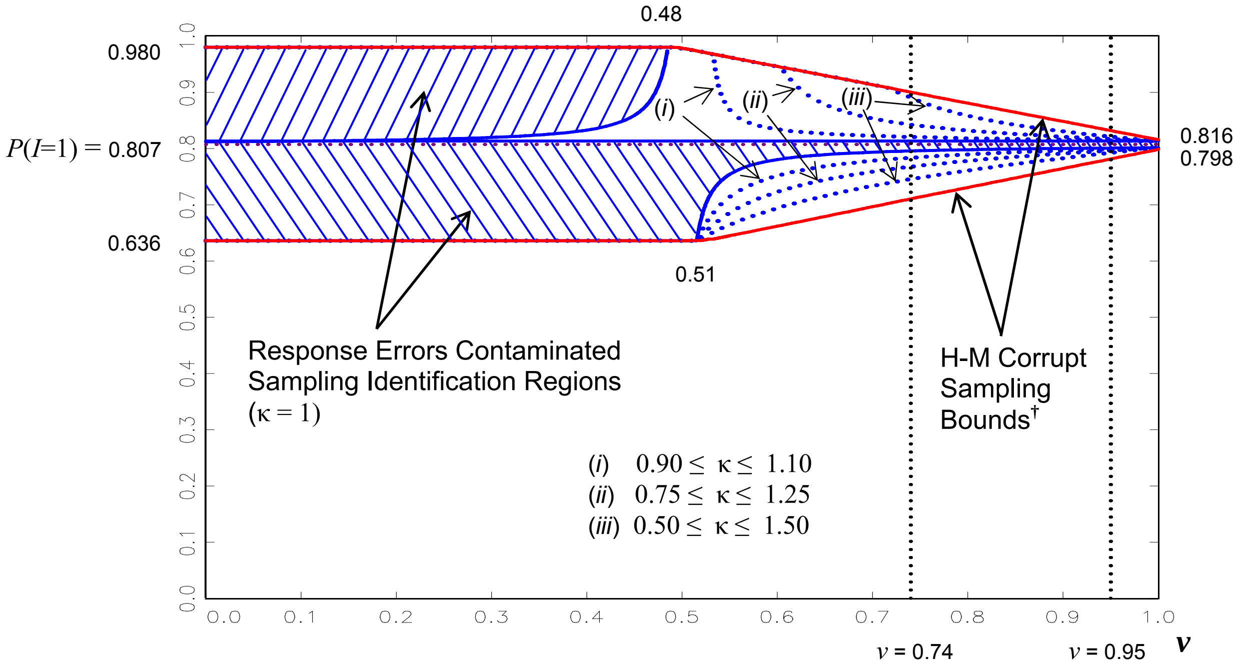


Figure 3. Identification Regions for the True Insured Rate Under "Partial Independence"



Notes: (a) v = minimum accurate reporting rate among unverified cases
(b) The reference values v=0.74 and v=0.95 are taken from Hill's (2006) analysis of the 1996 MEPS.
(c) The "partial independence" bounds in cases (*i*)-(*iii*) closely resemble the general shape of the H-M pure contaminated sampling bounds in Figure 2; i.e., the identification regions are not disjoint.
(d) The estimates reflect 5th percentile lower bounds and 95th percentile upper bounds.

Table 1

Identification Regions for the U.S. Nonelderly Population's True Insured Rate, P^o

v	H-M Corrupt Sampling Bounds	H-M Pure Contaminated Sampling Bounds	Proposition 1 Response Error Identification Regions
1.00	[0.807, 0.807]	[0.807, 0.807]	[0.807, 0.807]
	[0.798 0.816] [c]	[0.798 0.816]	[0.801 0.813]
0.95[a]	**[0.791, 0.824]**	**[0.799, 0.816]**	**[0.807, 0.807]**
	[0.782 0.832]	[0.789 0.824]	[0.801 0.813]
0.90	[0.774, 0.840]	[0.789, 0.825]	[0.806, 0.807]
	[0.765 0.848]	[0.778 0.834]	[0.800 0.813]
0.85	[0.758, 0.857]	[0.777, 0.836]	[0.805, 0.807]
	[0.748 0.865]	[0.767 0.844]	[0.799 0.813]
0.80	[0.742, 0.873]	[0.765, 0.847]	[0.804, 0.807]
	[0.731 0.881]	[0.753 0.856]	[0.798 0.813]
0.74[b]	**[0.722, 0.893]**	**[0.748, 0.864]**	**[0.802, 0.807]**
	[0.711 0.900]	[0.735 0.873]	[0.796 0.813]
0.70	[0.709, 0.906]	[0.735, 0.876]	[0.800, 0.807]
	[0.697 0.914]	[0.721 0.886]	[0.794 0.813]
0.60	[0.676, 0.939]	[0.694, 0.914]	[0.787, 0.807]
	[0.662 0.946]	[0.679 0.924]	[0.780 0.813]
0.50	[0.648, 0.972]	[0.648, 0.967]	[0.648, 0.807]
	[0.636 0.978]	[0.636 0.977]	[0.636 0.813]
0.40	[0.648, 0.977]	[0.648, 0.977]	[0.648, 0.807] ∪ [0.838, 0.977]
	[0.636 0.980]	[0.636 0.980]	[0.636 0.813] ∪ [0.833 0.980]
0.30	[0.648, 0.977]	[0.648, 0.977]	[0.648, 0.807] ∪ [0.825, 0.977]
	[0.636 0.980]	[0.636 0.980]	[0.636 0.813] ∪ [0.820 0.980]
0.20	[0.648, 0.977]	[0.648, 0.977]	[0.648, 0.807] ∪ [0.821, 0.977]
	[0.636 0.980]	[0.636 0.980]	[0.636 0.813] ∪ [0.815 0.980]
0.10	[0.648, 0.977]	[0.648, 0.977]	[0.648, 0.807] ∪ [0.819, 0.977]
	[0.636 0.980]	[0.636 0.980]	[0.636 0.813] ∪ [0.813 0.980]
0.00	[0.648, 0.977]	[0.648, 0.977]	[0.648, 0.807] ∪ [0.818, 0.977]
	[0.636 0.980]	[0.636 0.980]	[0.636 0.813] ∪ [0.812 0.980]

Notes: (a) v = minimum accurate reporting rate among unverified cases
(b) The reference values v=0.74 and v=0.95 are taken from Hill's (2006) analysis of the 1996 MEPS.
(c) 5th and 95th percentile bounds