

(forthcoming, *Review of Economics and Statistics*)

REGRESSION COEFFICIENT IDENTIFICATION DECAY IN THE PRESENCE OF
INFREQUENT CLASSIFICATION ERRORS

Brent Kreider*

Department of Economics
Iowa State University
bkreider@iastate.edu

Abstract—Recent evidence from Bound et al. (2001) and Black et al. (2003) suggests that reporting errors in survey data routinely violate all of the classical measurement error assumptions. The econometrics literature has not considered the consequences of fully arbitrary measurement error for identification of regression coefficients. This paper highlights the severity of the identification problem given the presence of even infrequent arbitrary errors in a binary regressor. In the empirical component, health insurance misclassification rates of less than 1.3 percent generate double-digit percentage point ranges of uncertainty about the variable's true marginal effect on the use of health services.

Keywords: Nonclassical measurement error, health insurance, corrupt sampling, binary regressor, classification error

JEL classification: C10, C20, I10

*Department of Economics, Iowa State University, Ames, IA 50011, 515-294-6237 (ph),
515-294-0221 (fax)

I received valuable comments from Chris Bollinger, Helle Bunzel, Harvey Lapan, Francesca Molinari, GianCarlo Moschini, Debasri Mukherjee, John Pepper, Justin Tobias, Quinn Weninger, Alex Zhylyevskyy, two anonymous referees, and seminar participants at Georgia State, Iowa State, the W.E. Upjohn Institute, and meetings of the Econometric Society. I gratefully acknowledge financial support from the Robert Wood Johnson Foundation through the Economic Research Initiative on the Uninsured (ERIU) PO#3000370614. The data come from a project coauthored with Steven Hill (Agency for Healthcare Research and Quality), and he provided valuable assistance with some of the computations in this paper.

I. Introduction

Explanatory variables in econometric regressions are often measured with error, and researchers have long understood that even random error can lead to substantially biased parameter estimates. Moreover, an emerging body of evidence from validation data suggests that patterns of measurement error in survey data often markedly violate the classical measurement error assumption (e.g., Bound et al. 2001). The classical assumption, imposed in nearly all empirical work that accommodates the possibility of data errors, specifies that reporting errors are independent of the true value of the underlying variable, all other regression covariates, and the stochastic disturbance. The standard result is that the coefficient estimate on the mismeasured variable is biased toward zero (e.g., Griliches 1986).

These independence assumptions may follow naturally in some applications, such as when errors arise passively from imprecise measuring devices. In many social science applications, however, the independence assumptions are unlikely to hold, even as a good approximation. Validation studies consistently reveal large degrees of response error in survey data for a wide range of self-reports, even for relatively objective variables.¹ In an important survey of the causes and consequences of measurement error, Bound et al. (2001) provide compelling evidence that inferences are often driven largely by untenable independence assumptions on the error generating process. In the context of most survey data, they find little reason to believe that reporting errors tend to be uncorrelated with the truth or other respondent characteristics. Instead, they find that most assessments of the consequences of reporting error, and proposed methods for correcting the biases (e.g., instrumental variables), have imposed strong and

¹ Black et al. (2003), for example, find that more than a third of respondents to the U.S. Census claiming to hold a professional degree have no such degree, with widely varying patterns of false positives and false negatives across demographic groups.

“exceedingly convenient” assumptions about the nature of the errors (Bound et al. 2001, p. 3708).

I study partial identification of regression coefficients given the possibility of infrequent but arbitrary classification errors in a binary regressor. Many key explanatory variables in econometric analyses are dichotomous. Common examples include the receipt of public transfers, health insurance status, labor force participation, on-the-job training, disability status, and pension status. I focus on simple regressions of health care utilization on health insurance status and other covariates in cases where true coefficients are assumed to be point-identified in the absence of insurance classification error. Once some insurance reporting errors are allowed, the true parameters can only be bounded.

Measurement error in a binary regressor automatically violates the classical assumption, except in degenerate cases, because errors must be mean-reverting (e.g., Aigner 1973). What may not be fully appreciated, however, is that the extreme nature of the measurement error in a binary regressor can result in severe identification deterioration of regression coefficients in the presence of very few classification errors. For a binary regressor, measurement error implies that the variable’s true value must be the polar opposite of its reported value.² Evidence from a variety of sources suggests the likelihood of substantial misreporting of health insurance in popular survey datasets, with unknown consequences for inferences (see Kreider and Hill, forthcoming, for discussion). Health insurance reporting errors must be negatively correlated with true insurance status. Moreover, reporting errors are also likely to violate the

² In contrast, no such relationship holds for a continuous variable, like income, where neither the self-reported value nor the truth is likely to lie at (or near) an endpoint of the variable’s domain.

“nondifferential” error assumption that, conditional on true insurance status and the other covariates, insurance classification errors must be unrelated to the use of health services.³

The usual method for correcting for measurement error in an explanatory variable is instrumental variables (IV) estimation. Standard IV is not valid, however, when the underlying mismeasured variable is binary because the measurement error is mean-reverting. Nor is it generally valid in a nonlinear regression setting (Amemiya 1985). When the classical measurement error properties do not hold, the literature has developed remedies, and partial remedies, in special cases. For example, Black et al. (2000) identify regression parameters for the case that health insurance errors are negatively correlated with true insurance status. They retain the assumption, however, that measurement error is independent of other covariates and the regression disturbance.

The consequences for identification of a mismeasured binary regressor were first addressed by Aigner (1973) in the context of linear models, with extended analysis in Bollinger (1996) and Frazis and Loewenstein (2003).⁴ Each analysis assumes that classification errors are nondifferential. Recently, there has been much progress in developing generalized IV methods to handle nonclassical measurement error in nonlinear models. Mahajan (2006), for example, retains the assumption of nondifferential classification errors in a binary regressor, but he relaxes the assumption that measurement error is independent of other covariates in the regression. Hu (2008) generalizes the approach to the case of misclassification of a general discrete explanatory

³ This assumption is violated if using health care informs some respondents about their true insurance status or if use of services depends on perceived insurance status in addition to true status. Moreover, the nondifferential assumption rules out the possibility that misclassification rates are informative about outcomes through their correlation with other observed covariates. Low-income households may be more prone to misreport their health insurance status, for example, because they experience more transitions in and out of true insurance coverage. Also, better-educated respondents may be more likely to be insured and more likely to accurately answer survey questions.

⁴ Fuller (1987) provides a comprehensive discussion of the consequences of classical measurement error, and of standard remedies, in the context of linear models. Carroll et al. (1995) expand the discussion to cover nonlinear cases.

variable. Hu and Schennach (2008) study the identifying power of auxiliary information that some characteristic of the distribution of the observed regressor (e.g., the median or mode), conditional on the true regressor, is left unaffected by the presence of measurement error.

Despite these important advances, the literature has not considered the case of fully arbitrary measurement error in either linear or nonlinear regression models.⁵ In the next section, I study identification of regression coefficients in a linear probability framework when a binary regressor may be arbitrarily misclassified. In Section III, I use a simulation approach to identify worst-case bounds on regression coefficients for both linear and probit specifications. My approach is motivated by the work of Horowitz and Manski (1995) who study partial identification of a random variable’s marginal distribution in the presence of “corrupt” data. They allow for the possibility of measurement error in a variable without imposing any assumptions on the nature of the error (see also Molinari 2008).

II. Arbitrary Classification Error in a Linear Model

Consider a simple linear probability model

$$Y = \alpha + \beta X_1^* + \delta X_2 + \varepsilon \tag{1}$$

where Y is a binary outcome, X_1^* and X_2 are binary regressors of interest, and ε is a random disturbance that is uncorrelated with the regressors. For concreteness, let $Y = 1$ indicate the use

⁵ Kreider and Pepper (2007) derive sharp bounds on unknown conditional distributions when the conditioning variable may be arbitrarily mismeasured, but their results do not apply to regression coefficients.

of health services within a given period, let $X_1^* = 1$ indicate being insured, and let $X_2 = 1$ indicate living in a metropolitan statistical area (MSA).⁶

As a departure from the previous literature, suppose that X_1^* may be arbitrarily misclassified subject to a limit on the maximum degree of data corruption. Specifically, suppose that X_1^* is unobserved and its observed counterpart X_1 may contain up to m misclassifications in a sample of size n (with the other variables measured without error). Then the maximum degree of corruption can be expressed as $q \equiv m/n$. Among the three observed binary variables Y , X_1 , and X_2 , there are $2^3 = 8$ possible types of misreporters. I restrict attention to the case that the degree of corruption is small enough that $q < \min \{P(Y = j, X_1 = k, X_2 = \ell)\}$ for all combinations of j , k , and ℓ equal to 0 or 1.⁷ I also assume that the regressors maintain full rank for each possible version of the true regressor matrix.

We can identify conservative degrees of identification decay of β and δ as a function of q by (1) assessing how the least squares estimates $\hat{\beta}$ and $\hat{\delta}$ must be modified when m respondents of the same particular type $\{j, k, \ell\}$ are hypothetically known to have misreported, and then (2) taking worst-case results across the eight types of potential misreporters. The resulting bounds are optimistically narrow in that allowing a mixture of types to misreport expands the range of possibilities for departures of $\hat{\beta}$ and $\hat{\delta}$ from their baseline values at $q = 0$. In Section III, I allow misreporters to be different types.

⁶ Standard regularity conditions (e.g., full rank) are assumed to hold. Shortcomings of the linear probability specification are well known, but I focus on this model for now to obtain tractable analytic results.

⁷ This assumption ensures that all m potential misreporters may be of any particular type (e.g., all might use health services, misreport being insured, and live in a rural community). Since I focus on very small misclassification rates (e.g., $q = 0.02$), this constraint is unlikely to matter in practice.

For each misreporter type, I derive true values of β and δ under the scenario that the m misreporters had reported correctly. Let $\kappa_0 = 1$ for health care users who misreported being uninsured, $\kappa_0 = -1$ for users who misreported being insured, and $\kappa_0 = 0$ for nonusers. Next, let $\kappa_1 = 1$ for respondents who misreported being uninsured and $\kappa_1 = -1$ for respondents who misreported being insured. Finally, let $\kappa_2 = 1$ for respondents who lived in an MSA, with $\kappa_2 = 0$ otherwise. Then define set K to be the set of vectors $(\kappa_0, \kappa_1, \kappa_2) \in \mathbb{R}^3$ such that $(\kappa_0, \kappa_1, \kappa_2)$ takes on one of the following types of potential misreporters: $\{-1, -1, 1\}$, $\{-1, -1, 0\}$, $\{0, 1, 1\}$, $\{0, 1, 0\}$, $\{1, 1, 1\}$, $\{1, 1, 0\}$, $\{0, -1, 1\}$, or $\{0, -1, 0\}$. Let K_s ($s = 1, \dots, 8$) denote element s of K .

Let $\{p_0, \sigma_0^2\}$, $\{p_1, \sigma_1^2\}$, and $\{p_2, \sigma_2^2\}$ denote the mean and variance of Y , X_1 , and X_2 , respectively, and let c_{01} , c_{02} , and c_{12} denote the covariance between Y and X_1 , Y and X_2 , and X_1 and X_2 , respectively. All of these parameters are identified by the observed data. Then the true value of β as a function of $\{\kappa_0, \kappa_1, \kappa_2\}$ and q , after appropriately modifying the standard least squares formula (see Appendix), is given by

$$\begin{aligned} \beta((\kappa_0, \kappa_1, \kappa_2); q) = & \left\{ \sigma_2^2 c_{01} - c_{12} c_{02} \right. \\ & \left. + q \left(\kappa_0 \sigma_2^2 + \kappa_1 \left[(p_2 - \kappa_2) c_{02} - p_0 \sigma_2^2 \right] \right) \right\} \\ & / \left\{ \sigma_1^2 \sigma_2^2 - c_{12}^2 - 2q \kappa_1 \left[\sigma_2^2 \left(p_1 - \frac{1}{2} \right) - (p_2 - \kappa_2) c_{12} \right] \right. \\ & \left. - q^2 \left[\sigma_2^2 + (\kappa_2 - p_2)^2 \right] \right\} \end{aligned} \quad (2)$$

When $q = 0$ (no errors), Equation (2) reduces to $\beta = (\sigma_2^2 c_{01} - c_{12} c_{02}) / (\sigma_1^2 \sigma_2^2 - c_{12}^2)$, which reduces further to the familiar expression $\beta = Cov(Y, X_1) / Var(X_1)$ when $q = 0$ and $c_{12} = 0$.

Accounting for the eight types of potential misreporters, we know the true value of β can be at least as small as $\beta' \equiv \min_{K_s \in K} \{\beta(s; q)\}$ and at least as large as $\beta'' \equiv \max_{K_s \in K} \{\beta(s; q)\}$.⁸

The true value of δ as a function of $\{\kappa_0, \kappa_1, \kappa_2\}$ and q (see Appendix) is given by

$$\begin{aligned} \delta((\kappa_0, \kappa_1, \kappa_2); q) = & \left\{ \sigma_1^2 c_{02} - c_{01} c_{12} \right. \\ & - q \left[\kappa_1 (\kappa_2 - p_2) c_{01} - \kappa_1 (1 - 2p_1) c_{02} + (\kappa_0 - \kappa_1 p_0) c_{12} \right] \\ & \left. + q^2 \kappa_1 \left[(\kappa_0 - \kappa_1 p_0) (p_2 - \kappa_2) - c_{02} \kappa_1 \right] \right\} \\ & / \left\{ \sigma_1^2 \sigma_2^2 - c_{12}^2 - 2q \kappa_1 \left[\sigma_2^2 (p_1 - \frac{1}{2}) - (p_2 - \kappa_2) c_{12} \right] \right. \\ & \left. - q^2 \left[\sigma_2^2 + (\kappa_2 - p_2)^2 \right] \right\} \end{aligned} \quad (3)$$

which reduces to $\delta = (\sigma_1^2 c_{02} - c_{01} c_{12}) / (\sigma_1^2 \sigma_2^2 - c_{12}^2)$ when $q = 0$, and further to

$\delta = Cov(Y, X_2) / Var(X_2)$ when $q = 0$ and $c_{12} = 0$. The true value of δ can be at least as small as $\delta' \equiv \min_{K_s \in K} \{\delta(s; q)\}$ and at least as large as $\delta'' \equiv \max_{K_s \in K} \{\delta(s; q)\}$.

Figure 1 traces out values of β' and β'' when $q = 0.02$ for various possible observed trivariate distributions of $\{Y, X_1, X_2\}$. In the figure, I set $c_{01} = c_{02} = 0$ such that the researcher's estimated values of β and δ based on the data (i.e., setting $q = 0$) are always zero. Frame A traces out β' and β'' as a function of p_1 when $p_0 = p_2 = \frac{1}{2}$.⁹ For example, if $p_1 = 0.3$ and $c_{12} = 0.10$, then $\beta' = -0.065$ and $\beta'' = 0.065$, a 13.1 percentage point range of uncertainty. The

⁸ It can be shown with examples (available upon request) that the true value of β can lie outside of $[\beta', \beta'']$ when the misreporters are allowed to be of different types. In known examples, the differences are slight.

⁹ As a technical note, the curves depicted in the figures exclude values of p_1 that are logically incompatible with the selected values of p_0 , p_2 , and c_{12} . For a distribution of three dichotomous variables, not all correlation matrices are possible. Incompatible combinations of p_0 , p_1 , p_2 , and c_{12} are identified using a simple algorithm provided in Chaganty and Joe (2006, p. 199).

ratio of the degree of uncertainty about β to the degree of uncertainty about the degree of data corruption, $r \equiv (\beta'' - \beta')/q$, is 6.5. For any c_{12} , the interval $[\beta', \beta'']$ is narrowest at $p_1 = \frac{1}{2}$ and expands as p_1 departs from $\frac{1}{2}$; specifically, the width is inversely related to σ_1^2 . Also, the bounds expand with $|c_{12}|$.¹⁰ The smallest degree of identification uncertainty arises when $p_0 = p_1 = p_2 = \frac{1}{2}$ and $c_{12} = 0$. In this case, $[\beta', \beta''] = [-0.0401, 0.0401]$. In Frame B, p_0 and p_2 are set equal to $\frac{1}{4}$ instead of $\frac{1}{2}$. Just as the width of $[\beta', \beta'']$ varies inversely with σ_1^2 , it also varies inversely with σ_0^2 and σ_2^2 . Returning to the case that $p_1 = 0.3$ and $c_{12} = 0.10$, $[\beta', \beta''] = [-0.105, 0.102]$, a 21 percentage point range of uncertainty with $r = 10.4$.

Thus far, I have focused exclusively on identification uncertainty. Sampling variability adds a second layer of uncertainty for inference since the population bounds $[\beta', \beta'']$ must be estimated. For small n , the uncertainty arising from sampling variability may be sufficiently severe that small degrees of classification error impose relatively little additional uncertainty. As n gets large, identification uncertainty eventually dominates. These two types of uncertainty are disentangled for some reference cases in Table 1. I constructed datasets of size $n = 200, 1000$, and 10,000 for various values of p_1 , with $p_0 = p_2 = \frac{1}{2}$ and $c_{01} = c_{02} = c_{12} = 0$, such that the researcher's estimated $\hat{\beta}$ is zero in all cases. Estimates of the worst-case values β' and β'' are presented for $q = 0.02$ along with their 90% confidence intervals (CI).¹¹

¹⁰ Formal results are available upon request. Intuitively, if the variance of X_1 is small, then either there are few observations involving $X_1 = 1$ or $X_1 = 0$. Since misreporting might be concentrated within these few observations, the potential impact of errors on coefficients is large. Since errors in X_2 may systematically occur for a particular value of X_2 , larger $|c_{12}|$ results in greater uncertainty about the true values of β or δ .

¹¹ Results are identical if p_1 is held constant at 0.5 across Columns (1)-(3) and $\text{Corr}(X_1, X_2)$ varies from 0 to ± 0.4 to ± 0.8 across these columns. For the parameter values considered in this table, the analytic optimistic

Moving left to right across columns, the CIs around β expand since the coefficient becomes less precisely estimated as p_1 departs from $\frac{1}{2}$ or as $|c_{12}|$ becomes large. At the same time, $\hat{\beta}'$ and $\hat{\beta}''$ move further away from 0. Their CIs also expand, as do the I-M (Imbens and Manski, 2004) CIs that contain the true value of β when $q = 0.02$ with 90% probability. Thus, for cases where the widths of the identification bounds are largest, uncertainty about the true parameter is relatively large even in the absence of classification error.¹²

Nevertheless, the table reveals that identification uncertainty can grow with $|p_1 - \frac{1}{2}|$ or $|c_{12}|$ more rapidly than sampling variability uncertainty. In Column (1) with $p_1 = \frac{1}{2}$ and $n = 200$, uncertainty from sampling variability dominates uncertainty from potential misclassification: the width of the CI under fully accurate classifications (0.23) nearly matches the width of the I-M bounds with classification error (0.27). In Column (3) when p_1 is far from $\frac{1}{2}$ (0.1 or 0.9), the relative role of identification uncertainty becomes stronger. For $n = 200$, the I-M CI $[-0.299, 0.299]$ is 54% wider than the CI $[-0.194, 0.194]$ around $\hat{\beta}$, compared with only 16% wider when $p_1 = \frac{1}{2}$. When $n=10,000$, the I-M CI $[-0.160, 0.160]$ when $q = 0.02$ is more than five times wider than the CI $[-0.027, 0.027]$ around $\hat{\beta}$. In the latter case, the CIs around $\hat{\beta}$, $\hat{\beta}'$, and $\hat{\beta}''$ ($[-0.027, 0.027]$, $[-0.168, -0.105]$, and $[0.105, 0.168]$) do not even overlap. Thus, with a very small classification error rate, the estimate of β obtained from the researcher's observed data can be sufficiently far away from the estimate of β that would be obtained from the error-free dataset, were this dataset known, that the CIs for these two estimates would not share any values.

bounds obtained using Equation (2) that restrict attention to a common misreporter type are identical to the worst-case bounds that allow for any combination of misreporter types.

¹² I thank the editor for bringing this point to my attention.

III. Identification Analysis Using MEPS Data

In this section, I study regression coefficient identification decay by constructing a real-world “population” consisting of 311 adults in the 1996 Medical Expenditure Panel Survey (MEPS).¹³ This population is defined to be all single white men between the ages of 20-50 who reported no disability, no military experience, and exactly 12 years of schooling. Among these adults, 46% used medical services in 1996 ($Y = 1$) and 56% reported being insured ($X_1 = 1$). These respondents comprise a subset of the 13,190 adults included in Kreider and Hill’s (forthcoming) universal health insurance analysis. Using validation data, they “verify” X_1 as being accurate for 7594 of these adults. For the remaining observations, true insurance status is unobserved. The sole objective in choosing the subsample of 311 adults was to obtain a relatively homogeneous subsample of manageable size for conducting the identification analysis below. After selecting on the other characteristics, the age range was chosen because it produced a convenient round number of 200 unverified insurance responses.

The basic idea will be to estimate a simple regression of health care utilization on insurance status and other covariates, define the resulting coefficient estimates to be the true parameters of interest in the absence of insurance misreporting (similar to a Monte Carlo approach, except guided by actual data), and then study how different the true parameters might actually be if the model is otherwise correctly specified but we allow for the possibility that some small fraction of insurance self-reports are in error. In what follows, I focus on the linear regression specified in Equation (1) and an analogous probit specification. In each case, I include four control variables in addition to MSA status (mean $\mu = 0.74$): X_3 is income level ($\mu = \$22,100$), X_4 indicates

¹³ The MEPS data are produced by the U.S. Agency for Healthcare Research and Quality and are available at the AHRQ Data Center.

excellent self-reported health ($\mu = 0.39$), X_5 indicates fair/poor self-reported health ($\mu = 0.06$), and X_6 is age ($\mu = 30.2$).¹⁴

Suppose that true health insurance status may be misreported by at most m respondents of unknown identity. In general, the total number of different ways the observed sample could deviate from a sample in which insurance status is never misclassified is given by $\sum_{j=1}^m n!/[j!(n-j)!]$. The number of possible deviations rapidly explodes as m increases. By the time even 1% of the population of size 13,190 is allowed to misreport ($m = 132$), the number of possible sample deviations exceeds 10^{270} . Unless the researcher has information that precludes certain patterns of errors, a valid identification analysis requires us to allow for the possibility that any pattern could occur. To conduct a feasible analysis, I study identification decay in the population of 311 adults when insurance status may be misreported in up to four of the 200 unverified cases. This framework yields 66,018,451 different possible configurations of true insurance status in the sample, a manageable number. I run separate regressions for each possible case and record sharp lower and upper bounds for each regression coefficient.

Table 2 presents results for the probit and linear probability models in Frames A and B, respectively. I focus on the probit results, but the two cases are very similar.¹⁵ In the absence of misreporting, the probit marginal effect 0.143 indicates that insured adults in this population are 14.3 percentage points more likely to use health services than the uninsured. If up to four respondents misreported true insurance status, however, then the true marginal effect could lie anywhere within the range [0.093, 0.193]. That is, potential misclassification in just 1.3% of the

¹⁴ The excluded category is good/very good health.

¹⁵ Differences between the probit and OLS models are quickly dwarfed by the uncertainty introduced by allowing for a small degree of reporting error. Slightly narrower bounds in the linear model are consistent with a theme in Bound et al. (2001) that parameter estimates are likely to be more sensitive to measurement error in nonlinear models than in linear models.

data is sufficient to generate a 10 point range of uncertainty about the true impact of insurance: $r = 7.8$. Importantly, this 10 point range does not reflect any uncertainty due to sampling variability. As discussed above, uncertainty about insurance status also translates into uncertainty about the coefficients on the other covariates. For example, residing in an MSA decreases the probability of using health services by 9.3 points if the data are accurate. Given the possibility of four insurance reporting errors, however, residing in an MSA may decrease the probability of using health services anywhere between 8.4 and 10.2 percentage points.

Figure 2 provides the frequency distribution for the probit marginal effect of being insured on the use of health services for all possible configurations of four or fewer insurance reporting errors (for the case that all configurations of insurance reporting errors are equally likely to occur). This figure reveals that reporting errors among many different types of respondents (not just worst-cases) lead to large impacts on the marginal effects. We might consider a stronger assumption that false positives and false negatives are known to be equally distributed across $Y = 0$ and $Y = 1$ outcomes (see the figure). This assumption has substantial identifying power as the true marginal effect is constrained to lie within a 3.9 point range. This interval remains quite large, however, given the maintained assumption that nearly 99% of the respondents reported their insurance status accurately, and there is no sampling variability. Moreover, there is little reason to believe that false positive and false negative reporting errors are evenly distributed.

The preceding results were closely replicated when I repeated the analysis using the full population of 13,190 adults using a method that approximates the parameter bounds.¹⁶

¹⁶ As discussed above, this population is much too large to conduct a simultaneous search for worst-case misreporters. Nevertheless, optimistically narrow worst-case bounds on the coefficients can be computed by searching for large-impact reporting errors sequentially instead of simultaneously. Specifically, we can start by finding the observation for which reclassifying insurance status would lead to the smallest (or largest) coefficient for the variable of interest. Then, leaving that report reclassified, we can find the next observation for which a

Analogous results for a Tobit model of health expenditures paint a similar picture. These results are available upon request.

IV. Conclusion

The econometrics literature has not considered the consequences of fully arbitrary measurement error for identification of regression coefficients. This paper highlighted the potential severity of the identification problem given the presence of even infrequent arbitrary errors in a binary regressor. In a linear probability setting, the rate of identification decay is inversely related to the observed variance of the misclassified regressor and positively related to the collinearity between this regressor and another covariate measured without error. In simple examples involving very small maximum error rates (e.g., less than 2%), the coefficient estimate obtained from the researcher's observed data can be sufficiently far away from the estimate that would be obtained from the error-free dataset, were this dataset known, that standard confidence intervals for these two estimates would not share any values.

Using a probit model in the empirical application, health insurance misclassification rates of less than 1.3% generate double-digit percentage point ranges of uncertainty about the variable's true marginal effect on the use of health services (prior to accounting for sampling variability). The wide nature of the bounds is not driven exclusively by rare combinations of misreporter types; many types of combinations yield coefficient estimates that lie far from the truth.

reclassification makes the largest additional impact, and so on. For the sample of 311 respondents, the sequential search bounds are only slightly narrower than the full search bounds.

Bound et al. (2001) argue that researchers using survey data should take much more seriously the possibility of nonclassical measurement error. For most microdata analyses, they find little reason to believe that reporting errors satisfy any of the classical assumptions and suggest that the assumptions generally reflect “convenience rather than conviction.” Consistent with this concern, Black et al. (2003) find that errors in self-reported education in an earnings regression are not only mean-reverting but also correlated with other covariates and the disturbance term. They suggest that standard IV estimates may be “highly biased” in this environment. Given large degrees of uncertainty about coefficient estimates obtained using bounding methods alone, IV methods generalized to account for nonclassical measurement error (e.g., Hu and Schennach 2008) may prove useful in this context.

REFERENCES

- Aigner, Dennis J., "Regression with A Binary Independent Variable Subject to Errors of Observation," *Journal of Econometrics*, 1 (1973), 49-60.
- Amemiya, Yasuo, "Instrumental Variable Estimator for the Nonlinear Errors-In-Variables Model," *Journal of Econometrics*, 28:3 (1985), 273-289.
- Black, Dan, Mark Berger, and Frank Scott, "Bounding Parameter Estimates with Non-Classical Measurement Error," *Journal of the American Statistical Association*, 95:451 (2000), 739-748.
- Black, Dan, Seth Sanders, and Lowell Taylor, "Measurement of Higher Education in the Census and CPS," *Journal of the American Statistical Association*, 98:463 (2003), 545-554.
- Bollinger, Christopher R., "Bounding Mean Regressions When a Binary Regressor is Mismeasured," *Journal of Econometrics*, 73 (1996), 387-99.
- Bound, John, Charles Brown, and Nancy A. Mathiowetz, "Measurement Error in Survey Data," In James J. Heckman and Edward E. Leamer (Eds.), Handbook of Econometrics, 5:59 (2001), 3705-3843.
- Carroll, Raymond J., David Ruppert, and Leonard A. Stefanski, Measurement Error in Nonlinear Models, Chapman and Hall, London, *Biometrics*, 53:3 (1997), 1180-1181.
- Chaganty, N. Rao and Harry Joe, "Range of Correlation Matrices for Dependent Bernoulli Random Variables," *Biometrika*, 93:1 (2006), 197-206.
- Frazis, Harley and Mark A. Loewenstein, "Estimating Linear Regressions with Mismeasured, Possibly Endogenous, Binary Explanatory Variables," *Journal of Econometrics*, 117:1 (2003), 151-178.
- Fuller, Wayne A., Measurement Error Models, Wiley, New York, (1987).

- Griliches, Zvi, "Economic Data Issues," In Zvi Griliches and Michael D. Intriligator (Eds.), Handbook of Econometrics, 3, Amsterdam, North-Holland, (1986), 1466-1514.
- Horowitz, Joel L. and Charles F. Manski, "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica*, 63:2 (1995), 281-302.
- Hu, Yingyao and Susanne M. Schennach, "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76:1 (2008), 195-216.
- Hu, Yingyao, "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution," *Journal of Econometrics*, 144:1 (2008), 27-61.
- Imbens, Guido and Charles F. Manski, "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72:6 (2004), 1845-1857.
- Kreider, Brent, and Steven C. Hill. (forthcoming). "Partially Identifying Treatment Effects with an Application to Covering the Uninsured," *Journal of Human Resources*.
- Kreider, Brent, and John V. Pepper, "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors," *Journal of the American Statistical Association*, 102:478 (2007), 432-441.
- Mahajan, Aprajit, "Identification and Estimation of Regression Models with Misclassification," *Econometrica*, 74:3 (2006), 631-665.
- Molinari, Francesca, "Partial Identification of Probability Distributions with Misclassified Data," *Journal of Econometrics*, 144:1 (2008), 81-117.

APPENDIX

Let X^* be an $n \times 3$ matrix consisting of a column vector of ones, the column vector X_1^* , and the column vector X_2 . While X^* is unobserved, we can nevertheless identify the true coefficient vector as $[\alpha(s; q), \beta(s; q), \delta(s; q)]' = \left(X^{*'} X^* \right)_s^{-1} \left(X^{*'} Y \right)_s$ conditional on knowing that m individuals of type s misreported. Then worst case coefficients across the eight values of s serve to identify bounds on $[\alpha, \beta, \delta]$ for the case of common misreporter types. Specifically, suppose X_1 is corrupted with m classification errors of type $K_s \in K$ associated with the values $\kappa_0, \kappa_1, \kappa_2$. Letting \sum_i denote summation over individuals, we can write

$$\begin{aligned} \left(X^{*'} X^* \right)_s &= \begin{bmatrix} n & \sum_i X_{1i} + m\kappa_1 & \sum_i X_{2i} \\ \sum_i X_{1i} + m\kappa_1 & \sum_i X_{1i}^2 + m\kappa_1 & \sum_i X_{1i} X_{2i} + m\kappa_1 \kappa_2 \\ \sum_i X_{2i} & \sum_i X_{1i} X_{2i} + m\kappa_1 \kappa_2 & \sum_i X_{2i}^2 \end{bmatrix} \\ &= n \begin{bmatrix} 1 & p_1 + q\kappa_1 & p_2 \\ p_1 + q\kappa_1 & p_1 + q\kappa_1 & c_{12} + p_1 p_2 + q\kappa_1 \kappa_2 \\ p_2 & c_{12} + p_1 p_2 + q\kappa_1 \kappa_2 & p_2 \end{bmatrix} \end{aligned}$$

$$\text{and } \left(X^{*'} Y \right)_s = \begin{bmatrix} \sum_i Y_i \\ \sum_i X_{1i} Y_i + m\kappa_0 \\ \sum_i X_{2i} Y_i \end{bmatrix} = n \begin{bmatrix} p_0 \\ c_{01} + p_0 p_1 + q\kappa_0 \\ c_{02} + p_0 p_2 \end{bmatrix}.$$

The coefficients in Equations (2) and (3) are then obtained from $[\alpha(s; q), \beta(s; q), \delta(s; q)]'$

$$= \left(X^{*'} X^* \right)_s^{-1} \left(X^{*'} Y \right)_s.$$

TABLE 1. — SHARP BOUNDS ON β WITH CONFIDENCE INTERVALS (CI) WHEN $q = 0.02$ ($\leq 2\%$ MISREPORTING)

| | | Various p_1 with $p_0 = p_2 = 0.5$, $c_{01} = c_{02} = 0$, and $\text{Corr}(X_1, X_2) = 0^a$ | | | | | |
|---|-------------|--|-----------------|----------------------|-----------------|----------------------|-----------------|
| | | (1) | | (2) | | (3) | |
| | | $p_1 = 0.5$ | | $p_1 = 0.3$ or 0.7 | | $p_1 = 0.1$ or 0.9 | |
| | | $\hat{\beta}$ | | $\hat{\beta}$ | | $\hat{\beta}$ | |
| OLS point estimate ($q = 0$): | | 0.000 | | 0.000 | | 0.000 | |
| 90% CI ^b | n = 200: | [-0.116, 0.116] | | [-0.127, 0.127] | | [-0.194, 0.194] | |
| | n = 1000: | [-0.052, 0.052] | | [-0.057, 0.057] | | [-0.087, 0.087] | |
| | n = 10,000: | [-0.016, 0.016] | | [-0.018, 0.018] | | [-0.027, 0.027] | |
| Worst case bounds for $q = 0.02$: ^c | | $\hat{\beta}'$ | $\hat{\beta}''$ | $\hat{\beta}'$ | $\hat{\beta}''$ | $\hat{\beta}'$ | $\hat{\beta}''$ |
| | | -0.0401 | 0.0401 | -0.0497 | 0.0497 | -0.137 | 0.137 |
| 90% CI | n = 200: | [-0.157, 0.076] | [-0.076, 0.157] | [-0.179, 0.080] | [-0.080, 0.179] | [-0.344, 0.071] | [-0.071, 0.344] |
| | n = 1000: | [-0.092, 0.012] | [-0.012, 0.092] | [-0.108, 0.008] | [-0.008, 0.108] | [-0.229, -0.044] | [0.044, 0.229] |
| | n = 10,000: | [-0.057, -0.024] | [0.024, 0.057] | [-0.069, -0.031] | [0.031, 0.069] | [-0.168, -0.105] | [0.105, 0.168] |
| 90% I-M ^d CI | n = 200: | [-0.134, 0.134] | | [-0.153, 0.153] | | [-0.299, 0.299] | |
| | n = 1000: | [-0.081, 0.081] | | [-0.095, 0.095] | | [-0.209, 0.209] | |
| | n = 10,000: | [-0.053, 0.053] | | [-0.064, 0.064] | | [-0.160, 0.160] | |

^a Results in this table are identical if p_1 is held constant at 0.5 across Columns (1)-(3) and $\text{Corr}(X_1, X_2)$ varies from 0 to ± 0.4 to ± 0.8 across these columns.

^b All confidence intervals presented in this table are heteroskedasticity-robust.

^c The values $\hat{\beta}'$ and $\hat{\beta}''$ reflect point estimates obtained from the data configurations that produce the smallest and largest estimates of β , respectively, when $q = 0.02$. The confidence intervals are obtained from these worst-case configurations. For the parameter values considered in this table, these estimated worst-case bounds that allow for any combination of misreporter types are identical to the analytic optimistic bounds obtained using Equation (2) that restrict attention to a common misreporter type.

^d Imbens and Manski (2004) confidence intervals that cover the true value of β with 90% probability when $q = 0.02$

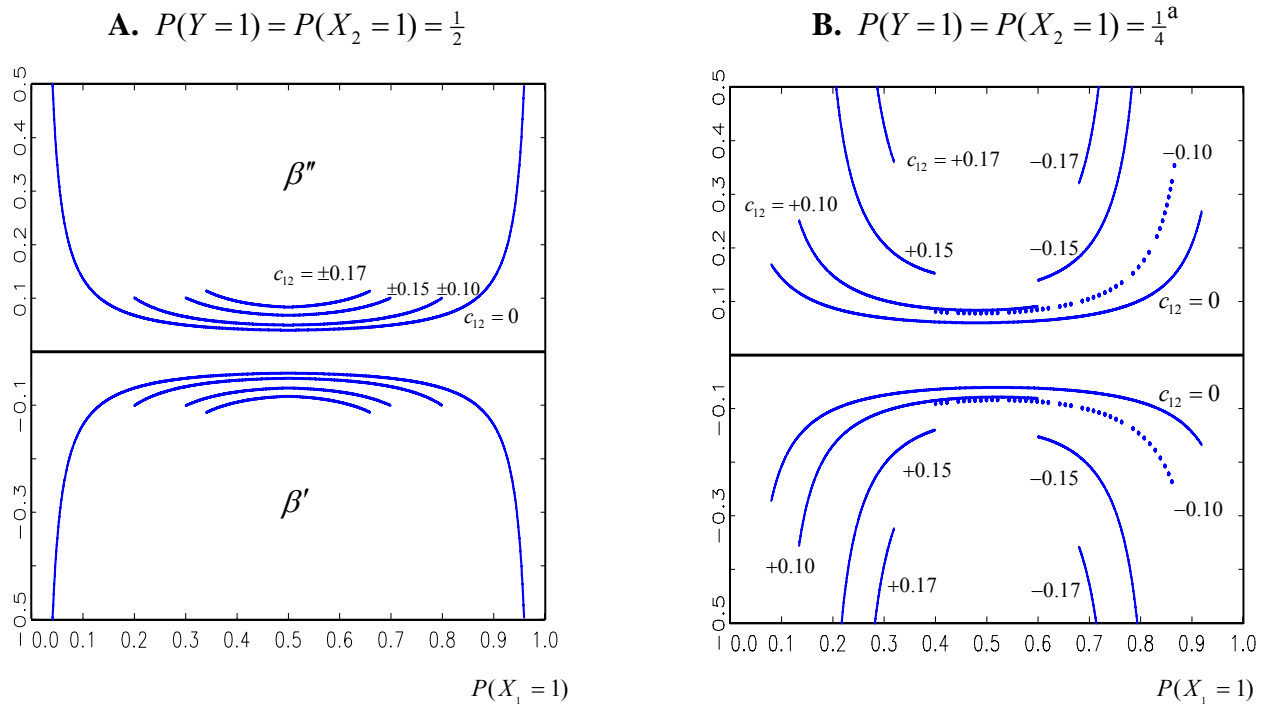
TABLE 2. — SHARP BOUNDS ON MARGINAL EFFECTS WHEN INSURANCE STATUS
MAY BE MISREPORTED BY UP TO 1.3 PERCENT OF THE POPULATION
(2 PERCENT OF THE UNVERIFIED POPULATION)

Dependent variable: $Y = 1$ if used health services in 1996

| | A. Probit Model | | B. Linear Probability Model (OLS) | |
|-------------------------------|------------------------|---|-----------------------------------|---|
| | No Errors ^a | Arbitrary Reporting Errors ^b | No Errors | Arbitrary Reporting Errors |
| | | LB | | UB |
| truly insured | 0.143 | [0.093 0.193] -35% +34% ^d width = 10 pts ^e , $r = 7.8^f$ | 0.134 | [0.090 0.177] -33% +32% width = 8.7 pts, $r = 6.8$ |
| resides in MSA | -0.0934 | [-0.102 -0.084] -10% +9% width = 1.8 pts, $r = 1.4$ | -0.0882 | [-0.095 -0.081] -8% +8% width = 1.5 pts, $r = 1.1$ |
| income/\$1,000 | 0.00308 | [0.0026 0.0035] | 0.00295 | [0.0025 0.0034] |
| excellent health ^c | -0.169 | [-0.179 -0.164] | -0.164 | [-0.171 -0.158] |
| fair/poor health | 0.330 | [0.306 0.353] | 0.323 | [0.288 0.352] |
| age | -0.00298 | [-0.0040 -0.0025] | -0.00284 | [-0.0036 -0.0023] |

Notes: ^aBaseline case that all health insurance classifications are accurate; ^bUp to four insurance classification errors allowed; ^cOmitted health category is “good/very good” health; ^dPercentage difference relative to no reporting errors case; ^eWidth of point estimate bounds in percentage points; ^fWidth of the bounds divided by q ($= 4/311 = 0.0128$), where q is the degree of potential data corruption

FIGURE 1. — VALUES OF β' AND β'' WHEN $\hat{\beta} = \hat{\delta} = 0$ AND $q = 0.02$



^aFrame B becomes its mirror image when $P(Y=1) = P(X_2=1)$ is set equal to $\frac{3}{4}$ instead of $\frac{1}{4}$.

FIGURE 2. — FREQUENCY DISTRIBUTION OF THE PROBIT MARGINAL EFFECT OF INSURANCE STATUS ON THE USE OF HEALTH SERVICES WHEN INSURANCE STATUS MAY BE MISREPORTED BY UP TO 1.3 PERCENT OF THE SAMPLE

