ELSEVIER

Journal of Econometrics 73 (1996) 387–399

JOURNAL OF Econometrics

Bounding mean regressions when a binary regressor is mismeasured

Christopher R. Bollinger

Policy Research Center, Georgia State University, Atlanta, GA 30303-3083, USA

(Received July 1992; final version received March 1995)

Abstract

In this paper I examine identification and estimation of mean regression models when a binary regressor is mismeasured. I prove that bounds for the model parameters are identified and provide simple estimators which are consistent and asymptotically normal. When stronger prior information about the probability of misclassification is available, the bounds can be made tighter. Again, a simple estimator for these cases is provided. All results apply to parametric and nonparametric models. The paper concludes with a short empirical example.

Key words: Measurement error; Binary variables; Identification
JEL classification: C10; C20

1. Introduction

The topic of measurement error has a long history in econometrics. In particular, it is well-known that when one or more regressors in a linear model are mismeasured, least squares estimation is generally not consistent. However, most of the literature has focused upon a continuous regressor. In this paper, I examine identification and estimation of bounds for the model parameters when the mismeasured regressor is a binary classification variable.

The model is formally represented by *Model I*:

$$Y = \alpha + \beta Z + u, \qquad \mathrm{E}[u \mid Z] = 0, \tag{1}$$

I thank Charles Manski, Arthur Goldberger, Mico Loretan, Chuck Michalopolous, Duncan Chaplin, an assistant editor, and two anonymous referees for many helpful comments and suggestions. Any errors are, of course, my responsibility.

SSDI 0304-4076(95)01730-2

$$\Pr[X = 0 \mid Z, Y] = (1 - p)(1 - Z) + qZ, \tag{2}$$

$$\Pr[X = 1 \mid Z, Y] = (1 - q)Z + p(1 - Z), \tag{3}$$

$$Z \sim \text{Bernoulli}(P_Z) \quad \text{with} \quad 0 < P_Z < 1, \tag{4}$$

$$p + q < 1. \tag{5}$$

The term p is the probability of reporting $X = 1$ when $Z = 0$, and q is the probability of reporting $X = 0$ when $Z = 1$. The researcher is interested in the parameters (α, β), but is only able to observe Y, the dependent variable, and X, the mismeasured version of Z. The focus in this paper will be on identification; hence the discussion focuses on population relationships.

Aigner (1973) showed that in this model, the least squares regression of Y on X does not yield consistent estimates of the parameters (α, β). The estimates will be asymptotically biased toward zero. Knowledge of p and q can be used to obtain consistent estimates (see Freeman, 1984; Aigner, 1973). Frisch (1934) showed that in a classical errors-in-variables model with one regressor (where the regressor is continuous and the error term is additive white noise), bounds for the parameters are identified. Klepper and Leamer (1984) generalized the result to the case where k continuous regressors are mismeasured. Klepper (1988) derives bounds for Model I with the additional assumption that $p = q < 0.5$. The assumption that $p = q$ is rather strong. Recent empirical evidence using validation data to estimate misclassification rates in survey data suggests that the assumption fails in practice (Freeman, 1984; Poterba and Summers, 1986; Mathiowetz and Duncan, 1988; Bollinger and David, 1993).

In Section 2, I derive bounds for β and the other parameters in Model I. Klepper's (1988) results cannot be obtained by simply substituting $p = q$ into the bounds derived here. In fact, the assumption $p = q$ is strong prior knowledge, and is fully utilized by Klepper, resulting in tighter bounds. In Section 3, I derive bounds when additional information, in the form of bounds on p and q, is available. In Section 4, I apply the results to bound the union wage differential. Proofs of all theorems and lemmas can be found in the Appendix.

2. Bounds for Model I

The discussion will focus on Model I, but the results can easily be extended to include linear models with other regressors which may be mismeasured also. The results can also be extended to nonparametric models with other correctly measured regressors.

The implications of assumption (5) and conditioning on Y in Eqs. (2) and (3) are worth examining. Assumption (5) insures that the misclassification is not so bad that X is independent of Z (the case where $p + q = 1$), or that the effective definition of the classification has been reversed (when $p + q > 1$) which would

occur if more than half the data were mismeasured. This also results in the covariance between X and Z being positive. An anonymous referee pointed out that a somewhat stronger assumption requiring $p \leqslant \frac{1}{2}$ and $q \leqslant \frac{1}{2}$ is reasonable. This assumption is not necessary for the result in Theorem 1, and can be imposed utilizing Theorems 2, 3, or 4 for various cases. Conditioning on Y in Eqs. (2) and (3) implies that the error process generating X is independent from the residual error, u, in the structural equation. If this assumption fails, the bounds derived here do not hold. Krasker and Pratt (1986) and Erikson (1993) study the case where the measurement error is not independent of the regression error in a classical errors-in-variables model.

It is helpful to establish some notation. Let b be the slope from the least squares projection of Y on X; let d be the inverse of the slope from the least squares projection of X on Y; let P_X be the marginal probability $X = 1$; let ρ_{XY}^2 be the squared correlation between X and Y. The variances of Y and X are represented by σ_Y^2 and σ_X^2, respectively. The covariance of Y and X is represented by σ_{XY}. Throughout this paper b, d, P_X, σ_Y^2, σ_X^2, and σ_{XY} are all observable.

From the assumptions of the model, restrictions on unobservable parameters are known. From (4) and (5), respectively, $0 < P_Z < 1$ and $p + q < 1$. By definition probabilities are nonnegative: $p \geqslant 0$ and $q \geqslant 0$. Variances are also nonnegative: $\sigma_u^2 \geqslant 0$ and $\sigma_Z^2 \geqslant 0$. These restrictions, combined with the first and second moments of the model and information about the error process from Eqs. (2) and (3) imply a set of constraints on the unknown structural parameters which give rise to Theorem 1. For the remainder of the paper I assume, without loss of generality, that $\beta \geqslant 0$, since it can be shown that $\operatorname{sign}(\beta) = \operatorname{sign}(b) = \operatorname{sign}(d)$, and if $\beta > 0$, then $0 < b < d$.

Theorem 1. Given Model I, if $\beta > 0$, then

$$0 < b \leqslant \beta \leqslant \max\{d \cdot P_X + b \cdot (1 - P_X), d \cdot (1 - P_X) + b \cdot P_X\}, \tag{6}$$

$$\mathrm{E}[Y] - d \cdot P_X \leqslant \alpha \leqslant \mathrm{E}[Y] - b \cdot P_X, \tag{7}$$

$$0 \leqslant p \leqslant P_X \cdot (1 - \rho_{XY}^2), \tag{8}$$

$$0 \leqslant q \leqslant (1 - P_X)(1 - \rho_{XY}^2). \tag{9}$$

If $\beta = 0$, then $b = \beta = 0$. These bounds utilize all information contained in the first and second moments of the observable data and are tight relative to this information.

Bounds are also available for P_Z and σ_u^2, but are not presented here. The lower bound on β was originally shown by Aigner (1973). I give an alternative proof and show that it is tight. The main focus of the paper will be on the upper bound.

The identification failure in any errors in variables model is due to the inability to differentiate between measurement error and the residual error term

u. The proof of Theorem 1 has two main parts. First, establish the maximum amount of measurement error which can feasibly be present in the system. Then, find the allocation of that error to the two distinct types of measurement error [errors of classifying 0's as 1's (represented by p) and errors of classifying 1's as 0's (represented by q)] which gives the largest feasible β.

To establish the maximum feasible amount of measurement error, first note that Model I can be rewritten as a classical errors-in-variables model. The mismeasured regressor is now $Z^* = \delta + \gamma Z$ (where $\delta = -p$ and $\gamma = 1 - p - q$), with $X = Z^* + \varepsilon$, where ε is uncorrelated with Z. The regression slope is now $\theta = \beta/\gamma$. It is well known that θ is bounded by b and d. The term θ is an index of the amount of measurement error in the system: for a given β, a larger θ implies larger p and/or q. The case where $\theta = d$ represents the maximal amount of measurement error.

Given the amount of measurement error, as indexed by θ, the allocation of this error to p and q is determined. From the classical errors in variables model, it can be shown that $V[\varepsilon] = V[X](1 - b/\theta)$. Here, the variance of ε can also be written as a function of p and q (see Lemma 3). These two equations can then be used to describe the set of feasible values of p and q given the amount of measurement error as indexed by θ. The largest feasible β over the set of feasible values for θ, p, and q can then be found.

If $P_X > \frac{1}{2}$, then the upper bound is associated with the case where $p = 0$ and $q = (1 - P_X)(1 - \rho_{XY}^2)$. If $P_X < \frac{1}{2}$, then the upper bound is associated with $q = 0$ and $p = P_X(1 - \rho_{XY}^2)$. Thus the upper bound is associated with a lopsided allocation of the total feasible measurement error. This lopsided error is misclassification from the largest of the two classes to the smallest. Further, the values of p and q which are associated with the upper bound on β are both less than $\frac{1}{2}$. Therefore, restricting p and q to be both less than $\frac{1}{2}$ would not alter the bounds for β.

An anonymous referee has pointed out that a parsimonious representation of the relationship between β, b, p, and q is

$$\beta = \frac{P_X(1 - P_X)(1 - p - q)}{(P_x - p)(1 - P_x - q)}. \tag{10}$$

Further, the referee remarks that the upper bound on β given that p and q are restricted to some set is the maximum of Eq. (10) on that set. The restriction that $p + q$ be less than one, or even that $p < \frac{1}{2}$ and $q < \frac{1}{2}$ is not sufficient to arrive at a bound. Even the result in Lemma 1 is not sufficient since p can be arbitrarily close to P_x and q can be arbitrarily close to $1 - P_x$. The result in Lemma 3 derives a more restrictive set for the feasible values of p and q utilizing the information in the variance of Y. Note that this set is a function of β. While it is possible to use the approach suggested by the referee, the approach here yields a simpler expression for the feasible values of p and q. Additionally, the

approach taken here highlights the fact that not only does the amount of error, as measured by the term θ, impact the bounds, but the allocation of that error to errors of omission or errors of commission is of critical importance also.

The bounds presented here are for a simple model with one regressor. Extending the bounds to apply to a linear model with other regressors is relatively straightforward for both the case where the other regressors are correctly measured, and the case where the other regressors may have classical measurement error. The details of that extension can be found in Bollinger (1993).

3. Imposing other information

In Section 2, I imposed the relatively weak assumption that $p + q < 1$. However, in many cases stronger information may be available. This information may take many forms. The cases I will discuss here are cases where there exist known M and K such that $p \leqslant M$ and $q \leqslant K$ or where there exist known m and k such that $p \geqslant m$ and $q \geqslant k$. Other cases are discussed in Bollinger (1993). The restrictions that $p \leqslant M$ and $q \leqslant K$ insure a stronger relation, that is less measurement error, between X and Z, and will only affect the upper bound on β since the lower bound on β is achieved when no measurement error is present. However, the restriction that $p \geqslant m$ and $q \geqslant k$ will affect both the upper and the lower bound on β.

Since the general case derived above implies that $p \leqslant P_X(1 - \rho_{XY}^2)$ and $q \leqslant (1 - P_X)(1 - \rho_{XY}^2)$ clearly any additional information about p and q must improve on at least one of these bounds. In particular, since the upper bound is associated with either the case where $p = 0$ and $q = (1 - P_X)(1 - \rho_{XY}^2)$ when $P_X > \frac{1}{2}$, or the case where $p = P_X(1 - \rho_{XY}^2)$ and $q = 0$ when $P_X < \frac{1}{2}$, the restriction that $p \leqslant M$ and $q \leqslant K$ must improve on the case associated with the general upper bound. Hence, if $P_X > \frac{1}{2}$, then K must be less than $(1 - P_X)(1 - \rho_{XY}^2)$; if $P_X < \frac{1}{2}$, then M must be less than $P_X(1 - \rho_{XY}^2)$.

If $p \leqslant M < P_X(1 - \rho_{XY}^2)$ and $q \leqslant K < (1 - P_X)(1 - \rho_{XY}^2)$, then two possible cases arise. In the first case, the original maximum feasible amount of measurement error is still feasible. Then the new information only affects the feasible allocations of the measurement error to p and q. In the second case, the values of M and K are so low that the original maximal amount of measurement error is no longer feasible. In this case, not only do the new bounds affect the feasible allocation of the measurement error, but the maximal feasible amount of measurement error is reduced.

Theorem 2 gives the upper bound for the case where $P_X < \frac{1}{2}$ and $p \leqslant M < P_X(1 - \rho_{XY}^2)$ and $K > (1 - P_X)(1 - \rho_{XY}^2)$. The case where $P_X > \frac{1}{2}$ is symmetric. Theorem 3 gives the upper bound for the case where $p \leqslant M < P_X(1 - \rho_{XY}^2)$ and $q \leqslant K < (1 - P_X)(1 - \rho_{XY}^2)$.

Theorem 2. Given Model I with $P_X < \frac{1}{2}$ *and the additional information that* $p \leqslant M < P_X(1-\rho_{XY}^2)$ *but* $K > (1-P_X)(1-\rho_{XY}^2)$ *for some known* M *and* K, *then*

$$\beta \leqslant \max\begin{Bmatrix} d\cdot P_X + b\cdot(1-P_X), \\ d(P_X - M) + b(1-P_X)\ (P_X/(P_X - M)) \end{Bmatrix}. \tag{11}$$

Theorem 3. Given Model I with the additional information that $p \leqslant M < P_X(1-\rho_{XY}^2)$ and $q \leqslant K < (1-P_X)(1-\rho_{XY}^2)$, *for some known* M *and* K, *then*

$$\beta \leqslant \max\begin{Bmatrix} d(1-P_X-K) + b\cdot P_X((1-P_X)/(1-P_X-K)) \\ d(P_X - M) + b(1-P_X)\ ((P_X/(P_X - M)) \end{Bmatrix}, \tag{12}$$

if

$$d \leqslant b\left[\frac{(1-P_X)P_X}{(1-P_X-K)(P_X-M)}\right]; \tag{13}$$

otherwise

$$\beta \leqslant (1-M-K)\, b\left[\frac{(1-P_X)P_X}{(1-P_X-K)(P_X-M)}\right]. \tag{14}$$

It may seem possible to 'bootstrap' up to tighter bounds by using Eqs. (8) and (9) from Theorem 1. Inspection of the results in Theorems 2 and 3 will demonstrate this approach will simply return the original upper bounds from Theorem 1.

The results for prior information bounding p and q from below are very similar. In this case, the lower bounds clearly rule out the minimum feasible amount of measurement error. Hence, the expression for the new minimum is similar to case two of Theorem 3. Since the upper bound on β occurs when either $p = 0$ or $q = 0$, the restrictions on p and q have an impact on the upper bound similar to case one of Theorem 3. The new bounds are given in Theorem 4.

Theorem 4. Given Model I with the additional information that $m \leqslant p$ *and* $k \leqslant q$, *for some known* m and k *and the condition that*

$$d \geqslant b\left[\frac{(1-P_X)P_X}{(1-P_X-k)(P_X-m)}\right], \tag{15}$$

then

$$\beta \geqslant (1-m-k)\, b\left[\frac{(1-P_X)P_X}{(1-P_X-k)(P_X-m)}\right], \tag{16}$$

and

$$\beta \leqslant \max\begin{Bmatrix} d(1-P_X-k) + b\cdot P_X((1-P_X)/(1-P_X-k)) \\ d(P_X - m) + b(1-P_X)\ (P_X/(P_X-m)) \end{Bmatrix}. \tag{17}$$

The condition in the theorem insures that the lower bounds do not rule out all feasible values for the measurement error. If p and q are bounded both from above and below, the upper bound on β is the least of the two upper bounds from Theorems 3 and 4.

4. Empirical example: Bounding the union wage differential

The simplest extension of the results above is to a linear structural equation where additional regressors are assumed to be measured without error. This also requires that the measurement error process for the mismeasured binary regressor must be independent of the other regressors. This case is represented by *Model II*:

$$Y = \alpha + \beta_1 Z_1 + \underline{\beta}_2' \underline{Z}_2 + u, \qquad \mathrm{E}[u \mid Z_1, \underline{Z}_2] = 0, \tag{18}$$

$$\Pr[X_1 = 0 \mid Z_1, \underline{Z}_2] = (1-p)(1-Z_1) + qZ_1, \tag{19}$$

$$\Pr[X_1 = 1 \mid Z_1, \underline{Z}_2, Y] = (1-q)Z_1 + p(1-Z_1), \tag{20}$$

$$Z_1 \sim \text{Bernoulli}(P_Z) \quad \text{with} \quad 0 < P_Z < 1, \tag{21}$$

$$p + q < 1. \tag{22}$$

Again, p and q are the misclassification rates. The researcher can observe Y, X_1, the mismeasured version of the binary regressor Z_1, and $\underline{Z}_2$, the vector of other correctly measured regressors. In the particular example here, Y is the natural log of average hourly earnings, Z_1 is the true union status (1 if a member of a union, 0 otherwise), while X_1 is the reported union status, and the vector $\underline{Z}_2$ contains the variables: Education in years, Potential Experience (Age − Education − 6), Potential Experience squared, Race (1 if black), and Gender (1 if female). This data set is a subsample from the May 1985 Current Population Survey (CPS) of size 533 from Berndt (1991). One observation was dropped since Age − Education − 6 was negative. I chose this data set for availability and reproducibility. A more comprehensive analysis utilizing more recent data can be found in Bollinger (1993). I assume that other variables are correctly measured and abstract from the problem of endogeneity of the Union variable in order to focus on the bounds derived here.

Bounds, similar to those derived in Section 2, can be shown for Model II. The linear projection of Y on $\underline{Z}_2$ yields the vector $\underline{b}_2$ as a biased (due to the omitted variable Z_1) estimate for $\underline{\beta}_2$ and an intercept term a. The linear projection of X_1 on $\underline{Z}_2$ yields slope coefficients $\underline{H}$, an intercept term h_0, and an r-squared of R^2_{XZ}. The residuals Y^* from the regression of Y on $\underline{Z}_2$ and the residuals X_1^* from the regression of X_1 on $\underline{Z}_2$ can be shown to have a structure almost identical to Model I. Hence, with only slight modification, the bounds from Model I can be applied to the residual model to bound β_1. The bounds on β_1 can then be used

with the biased coefficient estimates from the short regression of Y on $\underset{\sim}{Z}_2$ to obtain bounds on $\underset{\sim}{\beta}_2$. This result is summarized in Theorem 5. The term b is the slope from the projection of Y^* on X_1^*. The term d is the inverse of the slope from the projection of X_1^* on Y^*.

Theorem 5. Given Model II ($\beta_1 \geq 0$ without loss of generality),

$$b \leq \beta_1 \leq \max \begin{Bmatrix} d(P_X + (1 - P_X)R_{XZ}^2) + b(1 - P_X)\,(1 - R_{XZ}^2) \\ d((1 - P_X) + P_X R_{XZ}^2) + bP_X(1 - R_{XZ}^2) \end{Bmatrix}, \tag{23}$$

and the components of the slope vector $\underset{\sim}{\beta}_2$ are bounded by the terms $\underset{\sim}{b}_2 - \underset{\sim}{H}b$ and $\underset{\sim}{b}_2 - \underset{\sim}{H}d$. The lower and upper bound of each component are determined by the sign of each component in $\underset{\sim}{H}$. The intercept term is bounded by

$$\min\{a - bh_0, a - dh_0\} \tag{24}$$

and

$$\max\{a - bh_0, a - dh_0 + P_X(1 - R_{XZ}^2)\,(d - b)\}. \tag{25}$$

All of the terms in the bounds are easily estimable. In addition, estimable asymptotic variances can be derived using standard delta method results. The result can be modified when the other regressors in the model are potentially mismeasured as well. The bounds derived by Klepper and Leamer (1984) can be directly incorporated (see Bollinger, 1993).

Descriptive statistics for the sample are reported in Table 1. The estimated upper and lower bounds for the slope coefficients are reported in Table 2, with standard errors of the estimates in parentheses. Since the term $\underset{\sim}{b}_2 - \underset{\sim}{H}b$ is associated with the lower bound on β_1 and the term $\underset{\sim}{b}_2 - \underset{\sim}{H}d$ is associated with the upper bound on β_1, I have reported these as the vectors 'left' and 'right' respectively. The estimated bounds for p and q are reported in Table 3.

The 'right' bounds can be very large relative to the 'left' bounds. It is important to note that the 'left' bounds are also the estimates for the slope coefficients of the model if the measurement error is ignored. This implies that measurement error has the potential to cause significant bias. However, in many cases it is reasonable to bound p and q from above. I have chosen three sets of values for M and K (the upper bounds on p and q, respectively). The first case

Table 1
Descriptive statistics

	Ln Wage	Education	Experience	Black	Gender	Union
Mean	2.06	13.01	17.86	0.13	0.45	0.18
Std. error	0.53	2.61	12.37	0.33	0.50	0.38

Table 2
Estimated bounds for linear model

Variable	Left bounds	Right bounds
Union	0.21	5.48
	(0.05)	(1.23)
Constant	0.60	0.07
	(0.12)	(0.65)
Education	0.09	0.07
	(0.01)	(0.05)
Experience	0.04	− 0.01
	(0.01)	(0.03)
Experience squared	− 0.0005	-0.0001
	(0.0001)	(0.0007)
Black	− 0.12	− 0.70
	(0.05)	(0.35)
Gender	− 0.23	0.57
	(0.04)	(0.21)

Table 3
Estimated bounds for p and q

	Minimum	Maximum
p	0	0.1658
q	0	0.7546

sets $M = 0.13$ and $K = 0.20$. These values are chosen as representative of rather weak assumptions (relative to what is known from Table 3) to illustrate the sensitivity of the bounds to additional information. The second case sets $M = 0.1$ and $K = 0.1$. This value can be thought of as a 'folk theorem' in which measurement error is thought to be less than 10%. In the third case, I utilize results from Freeman (1984) and set $M = 0.023$ and $K = 0.081$. This case represents Freeman's (1984) worst-case estimates of misreporting union status in the CPS. The new 'right' bounds on all the parameters for each of these cases are reported in Table 4. Since the information utilized has no impact on the 'left' bounds, these remain the same as those reported in Table 2.

The most striking feature of the results presented in Table 4 is the sensitivity of the upper bound to additional information. Focusing only on the values for the union coefficient, one notes that even the first case with $M = 0.13$ and $K = 0.20$ results in substantial improvement in precision (width) of the bounds. The

Table 4
Estimated right bounds for linear model under stronger information

Variable	$M = 0.13$, $K = 0.20$	$M = K = 0.10$	$M = 0.023$, $K = 0.081$
Union	0.83	0.47	0.24
	(1.44)	(0.33)	(0.06)
Constant	0.52	0.57	0.60
	(0.22)	(0.13)	(0.13)
Education	0.09	0.09	0.09
	(0.01)	(0.01)	(0.01)
Experience	0.03	0.03	0.04
	(0.02)	(0.01)	(0.01)
Experience squared	− 0.0005	− 0.0005	− 0.0005
	(0.0002)	(0.0001)	(0.0001)
Black	− 0.21	− 0.15	− 0.13
	(0.20)	(0.07)	(0.05)
Gender	− 0.10	− 0.19	− 0.23
	(0.28)	(0.07)	(0.04)

estimated upper bound on the union coefficient falls from 5.48 to 0.83. In fact, it can be shown that the bound on p is, in this case, driving the result. Hence, the restriction on q could be relaxed even further with no degradation of the precision. As the additional information is strengthened, by setting $M = K = 0.1$, the bounds continue to tighten, but at a less dramatic rate: the new upper bound on the union coefficient falls to 0.47. Finally, when $M = 0.023$ and $K = 0.08$, the bounds fall to 0.24.

Mismeasurement may have great impact on the estimates for parameters in many models. The example here illustrates how serious this problem may be. One approach to gaining insight on the impact of measurement error is to utilize bounds, such as those presented here, to estimate the potential impact of measurement error.

Appendix

Proof of Theorem 1

This proof will focus on the upper bound. The lower bound is derived similarly.

Lemma 1. Given Model I, $P_X > p$ *and* $1 - P_X > q$.

Proof. By definition $P_X = P_Z(1-q) + (1-P_Z)p$. The result follows from the restrictions that $0 < P_Z < 1$ and $p + q < 1$. ■

Lemma 2. Given Model I, $\text{sign}(b) = \text{sign}(\beta) = \text{sign}(\theta)$, *and for* $\beta > 0$, $b \leqslant \theta \leqslant d$.

Proof. Model I can be rewritten as

$$Y = (\alpha - \delta\theta) + \theta Z^* + u, \tag{26}$$

$$X = Z^* + \varepsilon, \tag{27}$$

where $\gamma = 1 - p - q$, $\delta = -p$, $\theta = \beta/\gamma$, $Z^* = \delta + \gamma Z$, and ε is uncorrelated with Z and u. Then it can be shown that

$$\sigma_u^2 = \sigma_Y^2 - \theta\sigma_{XY}, \tag{28}$$

$$V[\varepsilon] = \sigma_X^2(1 - b/\theta), \tag{29}$$

$$\sigma_Z^2 = \sigma_X^2 b/\beta\gamma\,. \tag{30}$$

The first result in the lemma follows from the restrictions that $\sigma_Z^2 > 0$ and $p + q < 1$. The bound on θ follows from the restrictions that $\sigma_u^2 \geqslant 0$ and $V[\varepsilon] \geqslant 0$. ■

Lemma 3. Given Model I and any feasible value of θ *from Lemma 2,*

$$q = (1 - P_X)\left(1 - \left(\frac{P_X}{P_X - p}\right)\left(\frac{b}{\theta}\right)\right). \tag{31}$$

Proof. It can be shown that

$$V[\varepsilon] = \left(\frac{P_X - p}{1 - p - q}\right)q(1-q) + \left(\frac{1 - P_X - q}{1 - p - q}\right)p(1-p). \tag{32}$$

Setting this expression equal to Eq. (29) and solving for q yields the result. ■

Using the result of Lemma 3 and the definition of γ, the maximum feasible value for γ given θ is

$$\gamma^{\max} = \left\{1 - (1 - P_X)\left(1 - \frac{b}{\theta}\right), 1 - P_X\left(1 - \frac{b}{\theta}\right)\right\}. \tag{33}$$

By definition of θ, $\beta = \gamma * \theta$. Hence, the maximum value of β is

$$\beta^{\max} = \max_{\theta} \quad \gamma^{\max}\theta. \tag{34}$$

Solving (34) subject to the bounds on θ from Lemma 2 yields the upper bound on β. The lower bound is found by finding $\gamma^{\min}$ for given θ and similarly finding

$\beta^{\min}$. The bounds on α follow from

$$\alpha = \mathrm{E}[Y] - \theta(P_X - p) \tag{35}$$

and the results from Lemma 2 and Lemma 3. The bounds on p and q follow from Lemma 3 and the upper bound on θ. The bounds are tight since there exists p, q, P_Z, and $V[u]$ to support any β in the feasible region including the bounds themselves. **Q.E.D.**

Proof of Theorem 2

Theorem 2 gives a new upper bound when only p is restricted and $P_X < \frac{1}{2}$. If $P_X < \frac{1}{2}$, then the upper bound on β is achieved when $p = P_X(1 - \rho_{XY}^2)$. The restriction on p rules out this allocation. Specifically, the maximum value of γ used in the proof of Theorem 1 is no longer feasible. The maximum value of γ under the conditions in Theorem 2 occurs at either $p = M$ or $p = 0$. The proof is completed as above by noting that $\gamma^{\max}\theta$ gives maximum feasible β for a given θ, and $\theta = d$ gives the global maximum of β. **Q.E.D.**

Proof of Theorem 3

Lemma 4. Given Model I and the restrictions that $p \leqslant M < P_X(1 - \rho_{XY}^2)$ *and* $q \leqslant K < (1 - P_X)(1 - \rho_{XY}^2)$, *then*

$$\theta \leqslant \min\left\{d, b\left[\frac{(1 - P_X)P_X}{(1 - P_X - K)(P_X - M)}\right]\right\}. \tag{36}$$

Proof. From Eq. (31), for a particular value of θ to be feasible given the restrictions on p and q, it must be that

$$K \geqslant (1 - P_X)\left(1 - \left(\frac{P_X}{P_X - M}\right)\left(\frac{b}{\theta}\right)\right). \tag{37}$$

Rearranging this restriction and including the result from Lemma 2 gives the result. ■

Then the maximum value of γ occurs at either $p = M$ or $q = K$ or both. Complete the proof by evaluating $\gamma^{\max}\theta$ at the maximum value of θ. **Q.E.D.**

Proof of Theorem 4

Lower Bound: As in Lemma 4, for a particular value of θ to be feasible,

$$k \leqslant (1 - P_X)\left(1 - \left(\frac{P_X}{P_X - m}\right)\left(\frac{b}{\theta}\right)\right). \tag{38}$$

Rearranging gives the minimum value of θ feasible for given k and m. This value is only achieved when $p = m$ and $q = k$. The lower bound follows.

Upper Bound: As in Theorem 3, the maximum feasible value for γ given θ occurs at either $p = m$ or $q = k$, the feasible minimums of p and q. Complete the proof by evaluating $\gamma^{\max}\theta$ at the maximum value of $\theta = d$. **Q.E.D.**

Proof of Theorem 5

By definition $Y^* = \beta_1 Z^* + u$, where Z^* is the residual from the regression of Z_1 on $\underline{Z}_2$. Then,

$$V[X_1^*] = V[X_1](1 - R_{XZ}^2). \tag{39}$$

The results from Lemma 2 and Lemma 3 can be applied, yielding bounds on $\theta = \beta_1/\gamma$ and

$$q = (1 - P_X)\left(1 - \left(\frac{P_X}{P_X - p}\right)\left(\left(\frac{b}{\theta}\right)(1 - R_{XZ}^2) + R_{XZ}^2\right)\right). \tag{40}$$

Then, the upper and lower bounds for β_1 can be found as in Theorem 1.

By definition $\underline{b}_2 = \underline{\beta}_2 + \underline{F}\beta_1$, where $\underline{F}$ is the slope from the regression of Z_1 on $\underline{Z}_2$. Also, $\underline{H} = \gamma\underline{F}$. Rearranging and utilizing the bound on θ yield the bound on $\underline{\beta}_2$. **Q.E.D.**

References

Aigner, Dennis J., 1973, Regression with a binary independent variable subject to errors of observation, Journal of Econometrics 1, 49–60.

Berndt, Ernst R., 1991, The practice of econometrics: Classical and contemporary (Addison-Wesley, Reading, MA).

Bollinger, Christopher R., 1993, Measurement error in binary regressor with an application bounding the union wage differential, Ph.D. dissertation (University of Wisconsin, Madison, WI).

Bollinger, Christopher R. and Martin H. David, 1993, Modeling food stamp participation in the presence of reporting errors, Social Science Research Institute working paper no. 9310 (University of Wisconsin, Madison, WI).

Erikson, Timothy, 1993, Restricting regression slopes in the errors-in-variables model by bounding the error correlation, Econometrica 61, 959–970.

Frisch, R., 1934, Statistical confluence analysis by means of complete regression systems (University Institute for Economics, Oslo).

Freeman, Richard B., 1984, Longitudinal analysis of the effects of trade unions, Journal of Labor Economics 2, 1–26.

Klepper, Steven and Edward E. Leamer, 1984, Consistent sets of estimates for regressions with errors in all variables, Econometrica 52, 163–183.

Klepper, Steven, 1988, Bounding the effects of measurement error in regressions involving dichotomous variables, Journal of Econometrics 37, 343–359.

Krasker, William S. and John W. Pratt, 1986, Bounding the effects of proxy variables on regression coefficients, Econometrica 54, 641–655.

Mathiowetz, Nancy A. and Greg J. Duncan, 1988, Out of work, out of mind: Response error in retrospective reports of unemployment, Journal of Business and Economic Statistics 6, 221–229.

Poterba, James M. and Lawrence H. Summers, 1986, Reporting errors and labor market dynamics, Econometrica 6, 221–229.