Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Journal of Econometrics 133 (2006) 51–70
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

Bounding parameters in a linear regression model with a mismeasured regressor using additional information

Yingyao Hu*

Department of Economics, University of Texas at Austin, 1 University Station C3100, BRB 1.116, Austin, TX 78712, USA

Accepted 21 March 2005
Available online 17 May 2005

Abstract

This paper discusses a linear regression model with a mismeasured regressor in which the measurement error is correlated with both the latent variable and the regression error. We use a linear structure to capture the correlation between the measurement error and the latent variable. This paper shows that the variance of the latent variable is very useful for revealing information on the parameters which otherwise cannot be obtained with such a nonclassical measurement error. The main result is that the finite bounds on the parameters can be found using the variance of the latent variable, regardless of how severely the measurement error and the regression error are correlated, if the mismeasured regressor contains enough information on the latent one. This paper also discusses the special but interesting case of the latent variable being dichotomous. In this case, the mean of the latent variable may even reveal information on the correlation between the measurement error and the regression error. All the bounds developed in the paper are tight.

JEL classification: C10; C20

Keywords: Nonclassical measurement error; Dichotomous latent regressor

*Tel.: +1 512 475 8556; fax: +1 512 471 3510.
E-mail address: hu@eco.utexas.edu.
URL: http://www.eco.utexas.edu/~hu/.

0304-4076/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2005.03.009

1. Introduction

The measurement error model has increasingly been a topic of interest among researchers who want to estimate economic parameters such as the return to schooling and the union wage differential. When a regressor is mismeasured in a linear regression model, the least-squares estimator is generally not consistent, but at least some information can be inferred about the true parameters from the inconsistent estimators. These types of results are in the form of bounds on the parameters, which will hold asymptotically. Under the classical assumption that the measurement error is independent of the latent regressor and the regression error, it is well known that the regressions of x on y and y on x provide asymptotic bounds on the coefficient on x in the one-regressor case (Gini, 1921). However, the problem is more complicated in a multi-regressor context, and the existence of bounds is limited to certain cases. The classical result in the area is due to Koopmans (1937), who shows that such a generalization is possible only under very restrictive conditions. Patefield (1981) and Klepper and Leamer (1984) present a similar result. When further information on the measurement error distribution, such as bounds on the error variance, is available, narrower bounds on the parameters can be found (Bekker et al., 1984). Similar types of bounds are also discussed in Leamer (1982, 1987) and Klepper (1988b).

While the classical measurement error has been studied intensively, nonclassical measurement error has drawn more and more attention from researchers in recent decades. Bekker et al. (1987) discuss the case of errors in regressors and the regression error being correlated. Iwata (1992) and Krasker and Pratt (1986, 1987) show that bounds on these correlations may help find bounds on parameters of interest. Erickson (1993) provides a neat result when the measurement error is independent of the latent regressor but correlated with the regression error. As for empirical evidence of the nonclassical measurement error, Rodgers et al. (1993) suggest that the measurement error may be correlated with the latent variable. Bound et al. (2001) also find that the assumption that the measurement error is independent of the latent variable is strong and often implausible.

This paper discusses a linear measurement error model in which the measurement error is correlated with both the latent variable and the regression error. Let y denote the dependent variable, x^* denote the latent regressor and w denote the row vector of the other regressors (excluding the constant). Let α, β and γ be the intercept, the regression coefficients of x^* and w respectively, where γ is a column vector with the same dimension as w. Let u stand for the regression error. The linear regression model is as follows:

$$y = \alpha + \beta x^* + w\gamma + u \tag{1}$$

with $\mathrm{E}(u|x^*, w) = 0$. The researcher observes another variable x together with y and w as the proxy of the latent variable x^*. A critical assumption in this paper is that the conditional mean of the measurement error $v = x - x^*$ is linear in the latent

regressor x^*. Then,

$$x = p + rx^* + \varepsilon, \tag{2}$$

where $E(\varepsilon|x^*, w) = 0$. Eq. (2) implies that the measurement error v may be correlated with the latent variable x^*, and that the observed variable x may also contain a systematic shift p.

The linear structure in Eq. (2) can be justified as follows: first, if v and x^* are joint-normally distributed and $E(x|x^*, w) = E(x|x^*)$, the conditional mean of v on x^* is a linear function of x^*. Second, when x and x^* are two 0–1 dichotomous variables, x and x^* also satisfy Eq. (2).

The linear structure in Eq. (2) allows the correlation between the measurement error and the latent variable. Such a correlation has received increasing attention in the literature, especially in studies relating to earnings and wages. For example, Angrist and Krueger (1999) compare the self-reported hourly wage in CPS with corresponding employers' records, and find that the variance of the log self-reported wage is 0.355 while that of the employer-reported wage is 0.430. The fact that the latter is larger than the former implies that the measurement error v must be correlated with the true value x^* if we assume employers' records are accurate. This is because the variance of the self-reported wage σ_{xx} would be larger than that of the employer-reported wage $\sigma_{x^*x^*}$ if the measurement error v were uncorrelated with the true wage x^*. Eq. (2) implies that the conditional mean of the measurement error v is linear in x^*, i.e., $E(v|x^*, w) = p + (r-1)x^*$ and that $\sigma_{xx} \geqslant r^2\sigma_{x^*x^*}$. Therefore, we have $r \leqslant 0.91$ if we assume $r > 0$. The fact that $r < 1$ means that the measurement error in the self-reported wage is negatively correlated with the true wage. This is also consistent with the existing findings, such as those in Rodgers et al. (1993).

The method in Erickson (1993) is not applicable to this framework because the latent regressor in this paper is correlated with its measurement error. It has been shown that no informative bounds on the parameters of interest exist when measurement error is correlated with both the latent regressor and the error of the regression (Krasker and Pratt, 1986; Bekker et al., 1987; Erickson, 1989). Therefore, additional information is needed to find the bounds on the parameters of interest. Since we may observe the latent variable from other sources, the additional information may be the variance of the latent variable. In other words, the researcher may observe y, x and w in one data set and x^* in another data set. This framework is reasonable for several applications. For example, wages are usually mismeasured in the survey data, while the administrative data may contain accurately measured wages.

Other useful additional information may include the bounds on the parameter r. It is plausible to assume the parameter r is bounded away from zero if x contains enough information on x^*. We may then assume there exists an m such that $r \geqslant m > 0$. One can show $r = \rho_{xx^*}\sqrt{\sigma_{xx}/\sigma_{x^*x^*}}$ where ρ_{xx^*} is the correlation coefficient between x and x^*. Since σ_{xx} and $\sigma_{x^*x^*}$ are identified, a lower bound on ρ_{xx^*} implies a lower bound on r. When x and x^* are two 0–1 dichotomous variables, the lower bound on r implies an upper bound on the total misclassification probability. We will show that

this information is very useful for finding informative bounds on the parameters of interest.

The paper is organized as follows: Section 2 derives the bounds for a single regressor linear model. Section 3 provides the main results of the paper. A linear model with a dichotomous latent regressor is discussed in Section 4 as an application. Section 5 concludes the paper. The appendix includes all the proofs.

2. The single regressor model: an illustration

We consider a single regressor model in this section to illustrate how to find the bounds on the parameters using the variance of the latent variable. The model is formally represented by *model I*:

$$y = \alpha + \beta x^* + u, \tag{3}$$

$$x = p + rx^* + \varepsilon, \tag{4}$$

$$\mathrm{E}(u|x^*) = \mathrm{E}(\varepsilon|x^*) = 0, \tag{5}$$

$$r > 0, \tag{6}$$

The assumption $r>0$ is not restrictive because one can always use $-x$ as the proxy of x^*. We will also assume $\rho_{xy}>0$. This assumption is not restrictive either, since one can multiply the regression equation by -1 and discuss the bounds on $-\beta$ instead of β. Define

$$b = \frac{\sigma_{xy}}{\sigma_{xx}}, \quad d = \frac{\sigma_{yy}}{\sigma_{xy}}, \quad \rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_{xx}\sigma_{yy}}}, \quad r_u = \sqrt{\frac{\sigma_{xx}}{\sigma_{x^*x^*}}}. \tag{7}$$

The bounds on the parameter of interest β can be derived as follows.

Theorem 1. *Given model I, without loss of generality, assume* $\rho_{xy}>0$. *Then,*

1. $0<r\leqslant r_u$;
2. $-r_u\sqrt{b(d-b)}<\beta\leqslant r_u\sqrt{bd}$.

These bounds are tight. In the case that $\rho_{\varepsilon u}=0$, *we have*

1. $\rho_{xy}r_u\leqslant r\leqslant r_u$;
2. $r_u b\leqslant\beta\leqslant r_u\sqrt{bd}$.

Proof. See the appendix. □

Theorem 1 provides bounds on the regression coefficient β and also on the parameter r when the variance of the latent variable $\sigma_{x^*x^*}$ is known. The most important point here is that all these bounds are finite regardless of how severely the

measurement error and the regression error are correlated. Furthermore, these bounds are very easy to calculate.

The upper bound on β can be written as $\sqrt{\sigma_{yy}/\sigma_{x^*x^*}}$, which can be derived directly from the condition $\sigma_{uu} \geqslant 0$. This means the variance of βx^* is bounded by that of y. And the variance of x^* reveals β from the variation of βx^*. Therefore, we obtain a finite upper bound on β. If there are other regressors, say, w, it is possible that the variance of βx^* is larger than that of y because x^* and w can be either positively or negatively correlated. Therefore, we would expect that the upper bound on β is larger than $\sqrt{\sigma_{yy}/\sigma_{x^*x^*}}$ or is even infinite in a general linear regression model. This multivariate case will be discussed in the next section.

Theorem 1 shows that $\sigma_{x^*x^*}$ does help find the finite bounds on the parameters. The useful variation in σ_{xx} is $r^2\sigma_{x^*x^*}$. If r (or $\sigma_{x^*x^*}$) is very small given $\sigma_{x^*x^*}$ (or r), then the variable x does not contain much information on the latent variable. And the bounds on the parameters will be large in the absolute value. In the extreme case that $\sigma_{x^*x^*} \to 0$, the bounds on β approach infinity. That is the reason why we cannot find finite bounds on β if $\sigma_{x^*x^*}$ is unknown. In this simple linear regression model, $|\beta| \leqslant \sqrt{\sigma_{yy}/\sigma_{x^*x^*}}$ holds even when x does not contain any useful information, i.e., $r = 0$. That means β is always bounded. Unfortunately, this result can only be generalized in the unlikely case of the latent variable x^* being uncorrelated with other regressors.

To compare our result with the existing ones, suppose the variance of x^* is unknown. Define $x^e = \mathrm{E}(x|x^*)$. We can rewrite the model as

$$y = \alpha' + \theta x^e + u, \tag{8}$$

$$x = x^e + \varepsilon, \tag{9}$$

where $\alpha' = \alpha - p\beta/r$ and $\theta = \beta/r$. If ε and u are uncorrelated, it is well known that θ can be bounded by b and d. One can show that $\theta = \sigma_{xy}/\sigma_{x^ex^e}$ so that bounds on $\sigma_{x^ex^e}$ $(= r^2\sigma_{x^*x^*})$ can be found. But no further information on β and r can be revealed without knowing $\sigma_{x^*x^*}$. If ε and u are correlated, Erickson's results show the bounds on $\sigma_{x^ex^e}$ and θ. Again, no further information on β and r can be revealed if $\sigma_{x^*x^*}$ is unknown. With the variance of x^* known, we can reveal r from $\sigma_{x^ex^e}$. The bounds on r can be developed from those on $\sigma_{x^ex^e}$ and the bounds on β can also be derived from those on θ.

The information on the first moments helps bound the parameter α and p but does not help bound β and r in general. In some cases, however, we may have extra restrictions on p or r. These restrictions may make the first moments useful to bound β and r. This does happen when x and x^* are dichotomous. In the end, the bounds generated in this procedure are tight because there exist possible values of unobservables to support any β or r in the feasible region, including the bounds themselves. We will prove the tightness of the bounds in the next section, in which we consider a general linear regression model.

3. The general linear regression model

The single regressor model illustrates how the bounds can be developed with the variance of the latent variable known. This section shows that there are still certain features of the bounds which are not captured by this simple model. First of all, the bounds on β may be infinite because of the existence of other regressors. In order to find meaningful (or finite) bounds on β, we introduce a known lower bound on r, say, $r \geqslant m$. This assumption is meaningful because r should be bounded away from zero if x contains enough information on x^*. A similar assumption is also made in Bollinger (1996). Second, we will show that the bounds on β can be expressed as a function of r. Moreover, the upper bound function of β may not be monotonic in r. This makes it more complicated to find the bounds.

Consider a linear regression model named as *model II*:

$$y = \alpha + \beta x^* + w\gamma + u, \tag{10}$$

$$x = p + rx^* + \varepsilon, \tag{11}$$

$$\mathrm{E}(u|x^*, w) = \mathrm{E}(\varepsilon|x^*, w) = 0, \tag{12}$$

$$r > 0. \tag{13}$$

We are interested in the bounds on β, r and γ. Define $\sigma_{xx.w} = \sigma_{xx} - \Sigma_{xw}\Sigma_{ww}^{-1}\Sigma_{wx}$, $\sigma_{yy.w} = \sigma_{yy} - \Sigma_{yw}\Sigma_{ww}^{-1}\Sigma_{wy}$, $\sigma_{xy.w} = \sigma_{xy} - \Sigma_{xw}\Sigma_{ww}^{-1}\Sigma_{wy}$ and $\rho_{xy.w} = \sigma_{xy.w}/\sqrt{\sigma_{xx.w}\sigma_{yy.w}}$. Redefine $b = \sigma_{xy.w}/\sigma_{xx.w}$, $d = \sigma_{yy.w}/\sigma_{xy.w}$. The values b and d can be estimated directly from the observed data by regression of y on x, w and regression of x on y, w. Also define

$$r_l = \sqrt{\frac{\sigma_{xx} - \sigma_{xx.w}}{\sigma_{x^*x^*}}}, \quad r_m = \sqrt{\rho_{xy.w}^2 r_u^2 + (1 - \rho_{xy.w}^2) r_l^2}, \tag{14}$$

and the two bounds functions

$$\begin{aligned} \overline{\beta}(r) &= b(1 + R(r))r, \\ \underline{\beta}(r) &= b(1 - R(r))r, \end{aligned} \tag{15}$$

where

$$R(r) = \sqrt{\rho_{xy.w}^{-2} - 1}\sqrt{\frac{r_u^2 - r^2}{r^2 - r_l^2}}. \tag{16}$$

The bounds on β and r are developed from the functions $\overline{\beta}(r)$ and $\underline{\beta}(r)$. One can show that $\underline{\beta}(r)$ is always monotonic in r but $\overline{\beta}(r)$ may not be. The upper bound on β is found by maximizing $\overline{\beta}(r)$ with respect to r. The properties of the function $\overline{\beta}(r)$ depend on the value of $\rho_{xy.w}$ as follows.

Lemma 2. *There exists a ξ such that $\overline{\beta}(r)$ is monotone when $\rho_{xy.w} \leqslant \xi$, and has a unique local maximum and a unique local minimum on $r_l < r \leqslant r_u$ when $\rho_{xy.w} > \xi$.*

Proof. See the appendix. □

In fact, ξ is the value of $\rho_{xy.w}$ for which $\overline{\beta}(r)$ has a saddlepoint. The detailed derivation of ξ is in the proof of Lemma 2. We also define $r_{\max}$ as the solution of $\partial\overline{\beta}/\partial r = 0$ satisfying $\partial^2\overline{\beta}/\partial r^2 < 0$. The bounds on β and r can be found as follows.

Theorem 3. *Given model II, assume* $\rho_{xy.w} > 0$ *and* $r \geqslant m$. *Then*

1. $\max(r_l, m) \leqslant r \leqslant r_u$.

The bounds on β *are as follows*:

1. $m \leqslant r_l \Longrightarrow -\infty < \beta < \infty$;
2. $\rho_{xy.w} \leqslant \xi$ *and* $m > r_l \Longrightarrow \underline{\beta}(m) \leqslant \beta \leqslant \overline{\beta}(m)$;
3. $\rho_{xy.w} > \xi$ *and* $r_{\max} \geqslant m > r_l \Longrightarrow \underline{\beta}(m) \leqslant \beta \leqslant \max\{\overline{\beta}(m), \overline{\beta}(r_{\max})\}$;
4. $\rho_{xy.w} > \xi$ *and* $m > r_{\max} \Longrightarrow \underline{\beta}(m) \leqslant \beta \leqslant \overline{\beta}(m)$.

Furthermore, these bounds are tight. In the case that $\rho_{\varepsilon u} = 0$, *we have*

5. $r_m \leqslant r \leqslant r_u$;
6. $r_u b \leqslant \beta \leqslant r_m d$.

Proof. See the appendix. □

Theorem 3 considers a multivariate regression model, which is more applicable than that in Theorem 1. Theorem 3 requires both the variance of the latent variable and a meaningful lower bound on r to obtain bounds on β. The bounds on r are always finite even if m is unknown, while the bounds on β are finite only if $m > r_l$. The intuition of this result is that β can be finitely bounded as long as we know the mismeasured variable x contains enough correct information on x^*. These bounds are easy to compute because the bound functions are known and well defined.

Figs. 1 and 2 provide a straightforward way to demonstrate how these bounds are derived. In both figures, the horizontal axis stands for β and the vertical axis stands for r. The solid curve on the left is the graph of the lower bound function $\underline{\beta}(r)$ and the dotted curve on the right is the graph of the upper bound function $\overline{\beta}(r)$. The third curve in the figure is the graph of the implicit function $(r_u^2 - r_l^2)b - (r^2 - r_l^2)\beta/r = 0$. The area to the right of the third curve implies that β and r satisfy $(r_u^2 - r_l^2)b - (r^2 - r_l^2)\beta/r < 0$, i.e., $\rho_{\varepsilon u} < 0$. Since $\underline{\beta}(r)$ and $\overline{\beta}(r)$ are derived when $\rho_{\varepsilon u}^2 = 1$, the condition $\rho_{\varepsilon u}^2 \leqslant 1$ implies that all the possible combinations of β and r lie between $\underline{\beta}(r)$ and $\overline{\beta}(r)$ in the graph. Note that $\underline{\beta}(r) \to -\infty$ and $\overline{\beta}(r) \to \infty$ only when $r \to r_l$. Fig. 2 shows that the upper bound function $\overline{\beta}(r)$ is not monotonic if $\rho_{xy.w} > \xi$. But the bounds on β and r can still be found in this case by maximizing $\overline{\beta}(r)$ with respect to r.

From Theorem 3, we have $r_l \leqslant r \leqslant r_u$ in general. The lower bound is strictly positive if x^* is correlated with w. It implies that the other regressor w helps reveal certain information about r because w is uncorrelated with the error ε. Moreover, the bounds on β are not finite if and only if $r = r_l$. Therefore, we can find finite bounds

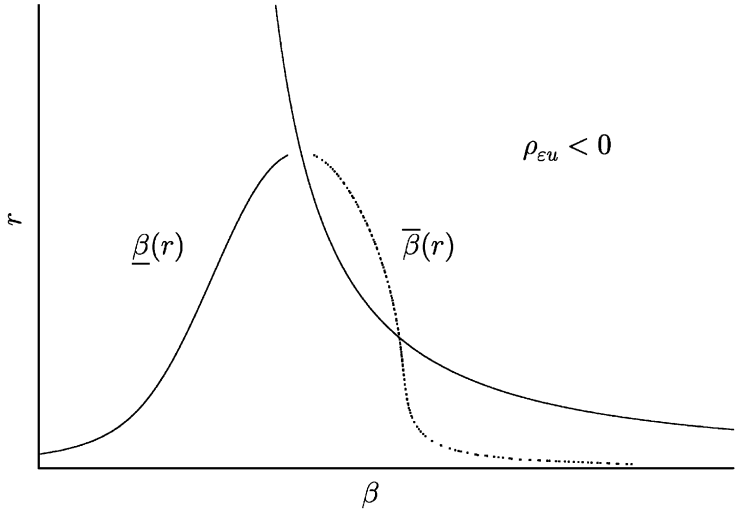

Fig. 1. The bound functions of β with $\rho_{xy.w} < \xi$.

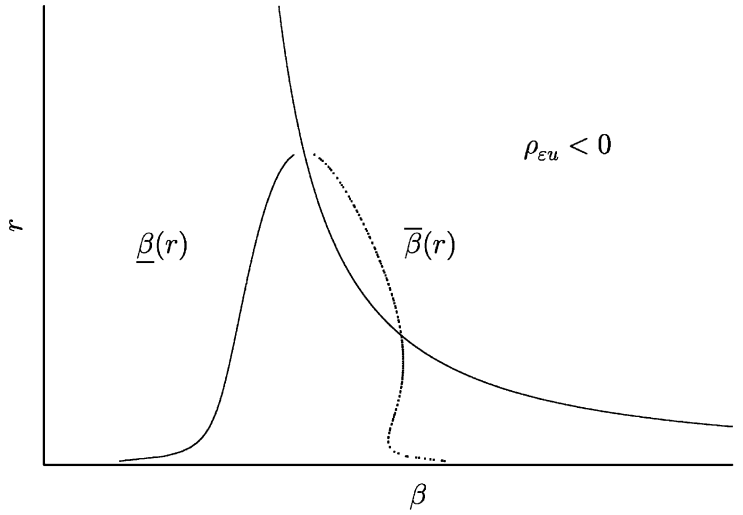

Fig. 2. The bound functions of β with $\rho_{xy.w} > \xi$.

on β for any value of $\rho^2_{\varepsilon u}$ if we know a lower bound m such that $r \geqslant m > r_l$. If there is no other covariate, we have $r_l = 0$, $m \to 0$, and $r_{\max} = r_m = \rho_{xy} r_u$. The results degenerate to those in theorem 1.

The bounds on the coefficient γ on the other regressor can be obtained through the following equation:

$$\gamma = \Sigma_{ww}^{-1}\Sigma_{wy} - \Sigma_{ww}^{-1}\Sigma_{wx}(\beta/r). \tag{17}$$

Therefore, it is enough to find the bounds on β/r. Since $\overline{\beta}(r)$ and $\underline{\beta}(r)$ are known functions, we can directly derive the bound functions of β/r to obtain those of γ. Let γ^j, q^j and s^j denote the jth elements of γ, $\Sigma_{ww}^{-1}\Sigma_{wy}$ and $\Sigma_{ww}^{-1}\Sigma_{wx}$, respectively. Then the jth equation in Eq. (17) can be written as

$$\gamma^j = q^j - s^j(\beta/r). \tag{18}$$

Without loss of generality, suppose $s^j \leqslant 0$. Then bounds on γ^j are summarized in the following theorem.

Theorem 4. *Given model II, assume* $\rho_{xy.w} > 0$, $s_j \leqslant 0$ *and* $r \geqslant m$. *Then*

1. $m \leqslant r_l \Longrightarrow -\infty < \gamma^j < +\infty$;
2. $m > r_l \Longrightarrow q^j - s^j\,\underline{\beta}(m)/m \leqslant \gamma^j \leqslant q^j - s^j\overline{\beta}(m)/m$.

These bounds are tight.

Proof. See the appendix. □

4. An application: the linear regression model with a dichotomous latent regressor

As mentioned before, one special case of the measurement error structure satisfying Eq. (2) is that x^* and x are 0–1 dichotomous variables with the following relationship:

$$P(x = 1|x^*, w) = (1 - q)x^* + p(1 - x^*). \tag{19}$$

The constants p and q are the misclassification probability. And $r = 1 - p - q$ in Eq. (2). The additional information here contains the mean of the latent dichotomous variable and a lower bound on r. Bollinger (1996) provides bounds on the parameters of interest based on the methodology developed by Klepper and Leamer (1984) and Klepper (1988a). Black et al. (2000) discuss the bounds on parameters when two noisy reports of the variable of interest are available. In these studies, the measurement errors are assumed to be independent of the dependent variable, conditional on regressors.

The difference between this model and the general linear regression model is that we have two extra restrictions: $1 \geqslant p \geqslant 0$ and $1 \geqslant q \geqslant 0$. These restrictions imply a new upper bound on r. Let $P_x = P(x = 1)$, $P_{x^*} = P(x^* = 1)$ and

$$r_d = \min\left\{\frac{1 - P_x}{1 - P_{x^*}}, \frac{P_x}{P_{x^*}}\right\}. \tag{20}$$

It will be shown that $r \leqslant r_d$ and $r_d \leqslant r_u$. That means the restrictions on p and q provide a smaller upper bound on r. This new upper bound on r may reveal information on $\rho_{\varepsilon u}$ as follows.

Lemma 5. *Given model I with Eq.* (19), *then*

1. $0 < r \leqslant r_d$;

2. $r_d > \rho_{xy} r_u \Longrightarrow 1 \geqslant \rho_{\varepsilon u} \geqslant -1$;
3. $r_d \leqslant \rho_{xy} r_u \Longrightarrow 1 \geqslant \rho_{\varepsilon u} \geqslant \left(1 - (1-\rho_{xy}^2)\frac{r_u^2}{r_u^2 - r_d^2}\right)^{1/2}$.

These bounds are tight.

Proof. See the appendix. □

Applying Theorem 1 to the dichotomous model with the additional restriction $r \leqslant r_d$, we have the bounds on β and r as follows.

Corollary 6. *Given model I with Eq.* (19), *then*

1. $r_d > \rho_{xy} r_u \Longrightarrow -r_u\sqrt{b(d-b)} < \beta \leqslant r_u\sqrt{bd}$;
2. $r_d \leqslant \rho_{xy} r_u \Longrightarrow -r_u\sqrt{b(d-b)} < \beta \leqslant \overline{\beta}(r_d)$.

These bounds are tight. In the case of $\rho_{\varepsilon u} = 0$, *assume* $r_d \geqslant \rho_{xy} r_u$. *Then*

1. $\rho_{xy} r_u \leqslant r \leqslant r_d$;
2. $r_u^2 b / r_d \leqslant \beta \leqslant r_u\sqrt{bd}$.

Proof. See the appendix. □

Bollinger (1996) shows the finite bounds on β when $\rho_{\varepsilon u} = 0$ and P_{x^*} is unknown. In the following discussion, we compare Bollinger's results with ours to show how P_{x^*} affects the bounds. It is straightforward to show that the lower bound developed in Corollary 6 is always larger than or equal to b, the lower bound in Bollinger's results, since $r_u^2 \geqslant r_d$. It takes two steps to show our upper bound $r_u\sqrt{bd}$ is no larger than the upper bound in Bollinger's results, i.e., $\max\{dP_x + b(1-P_x), d(1-P_x) + bP_x\}$. First, we need to find the bounds on P_{x^*} since P_{x^*} is unknown. Note that $\rho_{\varepsilon u} = 0$ implies $\sigma_{xy} = \beta r \sigma_{x^* x^*}$. As defined before, $\theta = \beta / r$. Then, θ is bounded by b and d. The bounds on P_{x^*} can be solved from

$$\sigma_{xy} = \theta r^2 P_{x^*}(1 - P_{x^*}),$$
$$b \leqslant \theta \leqslant d,$$
$$r \leqslant \min\left\{\frac{1-P_x}{1-P_{x^*}}, \frac{P_x}{P_{x^*}}\right\}$$

One can show that

$$1 - \frac{(1-P_x)^2}{(1-P_x)^2 + \sigma_{xy}/d} \leqslant P_{x^*} \leqslant \frac{P_x^2}{P_x^2 + \sigma_{xy}/d}. \tag{21}$$

Second, the upper bound on β can be solved by

$$\beta^{\max} = \max_{P_{x^*}} \sqrt{\frac{\sigma_{xx}}{P_{x^*}(1-P_{x^*})}}\sqrt{bd}. \tag{22}$$

The solution of maximization (22) subject to (21) is just the upper bound in Bollinger (1996), i.e., $\beta^{\max} = \max\{dP_x + b(1-P_x), d(1-P_x) + bP_x\}$. Therefore, the upper bound developed in this paper is smaller because P_{x^*} is known.

The mean of the latent regressor x^* plays two roles in the analysis. First, as shown in Eq. (22), P_{x^*} helps find narrower bounds because we do not need to maximize $\sqrt{\sigma_{xx}/\sigma_{x^*x^*}}\sqrt{bd}$ with respect to P_{x^*}. Second, P_{x^*} provides a new upper bound on r, i.e., $r \leqslant r_d$.

When there are covariates in the regression model with a latent dichotomous regressor, we have $r_l \leqslant r \leqslant r_u$ by Theorem 3. At the same time, $r \leqslant r_d$. The intersection of the two sets of r has to be nonempty for us to make the assumption that the mean of the latent variable is observed consistent with the data. Therefore, we must have

$$r_l \leqslant r_d. \tag{23}$$

The results in Lemma 5 and Corollary 6 can be extended to a general linear regression model with a dichotomous latent regressor by applying Theorem 3.

It is straightforward to check the above results with the real data. We consider the case of the regression error being uncorrelated with the misclassification error so that we can compare it with Bollinger's bounds. Take, for example, model I. Suppose the coefficient of x in the regression of y on x is 1, i.e., $b = 1$, and the reciprocal of the coefficient of y in the regression of x on y is 4, i.e., $d = 4$. Let $P_{x^*} = 0.3$, $p = 0.2$ and $q = 0.6$. The bounds on β suggested by this paper are $[1.06, 1.83]$, while Bollinger's bounds are $[1, 3.22]$. The result shows that the variance of the latent variable does help find narrower bounds.

It is also necessary to discuss when the informative (or finite) bounds can be found. As mentioned before, the bounds on β are finite if there exists an m such that $r \geqslant m > r_l$. The constraint $r \geqslant m$ means an upper bound of $p + q$, i.e., $p + q < 1 - m$. It suggests that the sum of misclassification probabilities is less than 1. Therefore, if we know the total misclassification probability $p + q$ is not very large, we may find informative bounds on β regardless of how severely the two errors are correlated. We consider the model II. Let b, d, P_{x^*}, p and q take the same values as shown above. Suppose $\sigma_{xx.w} = 0.19$.[1] Then, $r_l = 0.107$. Therefore, if we know an m such that $p + q \leqslant m < 0.893$, we can always get finite bounds on β. Suppose we know $m = 0.85$, then the bounds on β are $[-2.18, 2.48]$ for any $\rho_{\varepsilon u}$.

In reality, researchers can obtain estimates of $\sigma_{x^*x^*}$ and r from a validation sample. For example, we consider the estimation of wage function with college education as a 0–1 dichotomous independent variable. If an individual has at least some college education, this variable equals 1, otherwise it equals 0. The education level is subject to reporting error in most survey samples. Kane et al. (1999) use a validation sample containing self-reported data and transcript data to show the empirical joint distribution of true education level and misreported education level (Table 1). From this validation study, one can find that $\widehat{P}_{x^*} = 0.573$, $\widehat{p} = 0.1235$ and $\widehat{q} = 0.0681$ so that $\widehat{r}$ is equal to 0.8084. Therefore, one can bound the coefficient on the education variable in the wage function. Furthermore, one may find narrower bounds on the

[1]Note the inequality $\sigma_{xx} \geqslant \sigma_{xx.w} \geqslant \sigma_{xx} - r^2\sigma_{x^*x^*}$ must hold.

Table 1
Sample proportions of true and misreported education in Kane et al. (1999).

Transcript data	Self-reported data		
	No college education	College education	Total
No college education	0.376	0.053	0.429
College education	0.039	0.534	0.573
Total	0.415	0.587	

parameter β with other types of additional information on $p+q$ (or r). For example, let $p=q$ or $p=0$. Then, $r=(P_x-P_{x^*})/(1-2P_{x^*})$ or $r=P_x/P_{x^*}$. These conditions also provide narrower bounds on β (Table 1).

Another example of the use of a validation sample with a continuous regressor would be bounding the impact of earnings on consumption. When a consumption function is estimated using a survey sample, researchers usually are worried about the reporting error in earnings. Bound et al. (1994) use a validation sample containing 416 observations of self-reported earnings in PSID together with corresponding employers' records. They provide not only the sample variance of self-reported and true earnings, but also the sample correlation coefficient between self-reported and true earnings. Their study shows that the sample variance of true log earnings $\widehat{\sigma}_{x^*x^*}$ is 0.0416, that of self-reported log earnings $\widehat{\sigma}_{xx}$ is 0.0488, and the sample correlation coefficient between the two earnings is 0.8862. That means $\widehat{r}$ is equal to 0.9598, and therefore the 95% confidence interval of r is $[0.9115, 1.0082]$ based on its asymptotic distribution. One may then take 0.9115 as a lower bound on r. Thus, the bounds in this paper apply even if the reporting error and the regression error are correlated.

Bound et al. (2001) provide an excellent summary of these validation studies. Moreover, most of those studies provide the sample variance of the true value x^* and a point or interval estimate of r, which makes the method in this paper easy to apply.

5. Conclusion

This paper discusses a linear regression model with a mismeasured regressor under the assumption that the variance of the latent regressor is available. The main result is that the parameters of interest can be finitely bounded with additional information, the variance of the latent variable and an additional lower bound on the parameter r, regardless of how severely the measurement error is correlated with the regression error. If the regression error and the measurement error are uncorrelated, the variance of the latent regressor helps provide narrower bounds compared with those in the existing results. We also discuss the model with a latent dichotomous regressor as an application of the general result. In this case, the additional information needed includes the mean of the latent variable, and an upper bound on the total misclassification probability. The additional information may

lead to bounds not only on the parameters of interest, but also on the correlation coefficient between the measurement error and the regression error. The presented results suggest that the variance of the latent variable is very useful in solving the nonclassical measurement error problem in the linear regression model.

Acknowledgements

I would like to thank Geert Ridder, Robert Moffitt, Matt Shum, Randal Watson, an editor, and anonymous referees. All errors are mine.

Appendix

Proof (*Theorem 1*). By the assumptions in model I, we have

$$\sigma_{yy} = \beta^2 \sigma_{x^*x^*} + \sigma_{uu}, \tag{24}$$

$$\sigma_{xy} = \beta r \sigma_{xx^*} + \sigma_{\varepsilon u}, \tag{25}$$

$$\sigma_{xx} = r^2 \sigma_{x^*x^*} + \sigma_{\varepsilon\varepsilon}. \tag{26}$$

The sign of $\rho_{\varepsilon u}$ is the same as $\sigma_{\varepsilon u}$. Then,

$$sign(\rho_{\varepsilon u}) = sign(\sigma_{xy} - \beta r \sigma_{x^*x^*}). \tag{27}$$

The expression $\rho_{\varepsilon u} = \sigma_{\varepsilon u}/\sqrt{\sigma_{\varepsilon\varepsilon}\sigma_{uu}}$ [2] leads to

$$\rho_{\varepsilon u}^2(\sigma_{xx} - r^2\sigma_{x^*x^*})(\sigma_{yy} - \beta^2\sigma_{x^*x^*}) = (\sigma_{xy} - \beta r \sigma_{x^*x^*})^2. \tag{28}$$

Rearranging the terms and dividing by $\sigma_{x^*x^*}^2$, we have

$$A\beta^2 + B\beta + C = 0, \tag{29}$$

where

$$A = r^2 + \rho_{\varepsilon u}^2(r_u^2 - r^2),$$
$$B = -2br_u^2 r,$$
$$C = b^2 r_u^4 - \rho_{\varepsilon u}^2 b d r_u^2 (r_u^2 - r^2).$$

It can be shown that

$$B^2 - 4AC = 4bdr_u^2\rho_{\varepsilon u}^2(r_u^2 - r^2)[(1 - \rho_{\varepsilon u}^2)r^2 + (\rho_{\varepsilon u}^2 - \rho_{xy}^2)r_u^2].$$

The existence of β requires $B^2 - 4AC \geqslant 0$, which leads to the bounds on r for different values of $\rho_{\varepsilon u}$. The bounds on β are derived as follows: first, let

$$\overline{\beta}(r, \rho_{\varepsilon u}^2) = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \quad \text{and} \quad \underline{\beta}(r, \rho_{\varepsilon u}^2) = \frac{-B - \sqrt{B^2 - 4AC}}{2A}. \tag{30}$$

[2] Without loss of generality, we set $\rho_{\varepsilon u} = 0$ if $\sigma_{\varepsilon\varepsilon} = 0$ or $\sigma_{uu} = 0$.

Define $\overline{\beta}(r) = \overline{\beta}(r, 1)$ and $\underline{\beta}(r) = \underline{\beta}(r, 1)$. Since $\rho_{\varepsilon u}^2 \leqslant 1$, these two functions provide bounds on β, i.e., $\underline{\beta}(r) \leqslant \beta \leqslant \overline{\beta}(r)$. One can show that $\underline{\beta}(r)$ is increasing in r and $\overline{\beta}(r)$ reaches its unique maximum at $\rho_{xy} r_u$. Therefore, we have $\underline{\beta}(0) < \beta \leqslant \overline{\beta}(\rho_{xy} r_u)$. It is straightforward to show $\underline{\beta}(0) = -r_u\sqrt{b(d-b)}$ and $\overline{\beta}(\rho_{xy} r_u) = r_u\sqrt{bd}$.

In the case that $\rho_{\varepsilon u} = 0$, $\beta = \sigma_{xy}/r\sigma_{x^*x^*}$. And $\sigma_{uu} \geqslant 0$ implies $\sigma_{yy} \geqslant \beta^2 \sigma_{x^*x^*}$. These two conditions on β and r lead to $\rho_{xy} r_u \leqslant r$. The condition $r \leqslant r_u$ follows from $\sigma_{\varepsilon\varepsilon} \geqslant 0$. The bounds on β follow from the bounds on r.

Since Theorem 1 considers a special case of the model in Theorem 3, we will show the tightness of the bounds in the proof of Theorem 3. □

Proof (*Lemma 2*). We consider the upper bound function $\overline{\beta}(r) = b(1 + R(r))r$. The shape of the function $\overline{\beta}(r)$ depends on the value of $\rho_{xy.w}$ given r_l and r_u. We will show that there exists a ξ such that $\overline{\beta}(r)$ is monotone when $\rho_{xy.w} \leqslant \xi$, and that $\overline{\beta}(r)$ has a unique local maximum and a unique local minimum on $r_l < r \leqslant r_u$ when $\rho_{xy.w} > \xi$. In other words, if $\rho_{xy.w} < \xi$, then $\partial\overline{\beta}(r)/\partial r = 0$ has no roots on $r_l < r \leqslant r_u$; if $\rho_{xy.w} > \xi$ then the function $\partial\overline{\beta}(r)/\partial r = 0$ has two roots; if $\rho_{xy.w} = \xi$ then the function has only one root. From $\partial\overline{\beta}(r)/\partial r = 0$, we have

$$\frac{1}{\sqrt{\rho_{xy.w}^{-2} - 1}}\eta(r) = 1 + \frac{r_l^2}{r_u^2 - r_l^2}(\eta(r)^2 + 1)^2, \tag{31}$$

where

$$\eta(r) = \sqrt{\frac{r_u^2 - r^2}{r^2 - r_l^2}}. \tag{32}$$

Since $r_l < r \leqslant r_u$, we have $\eta \in [0, \infty)$ and $\partial\eta(r)/\partial r < 0$. The left-hand side of Eq. (31) is a linear function of η whose slope is a strictly increasing function of $\rho_{xy.w}$ ($\rho_{xy.w} > 0$). The right-hand side of Eq. (31) is a simple quartic function strictly increasing in η. Given r_u^2 and r_l^2, the right-hand-side function is fixed while the slope of the linear function on the left-hand side changes with $\rho_{xy.w}$. If $\rho_{xy.w}$ is close to 1, the linear function is so steep that it will intersect the function on the right-hand side. That means $\partial\overline{\beta}(r)/\partial r = 0$ has two roots. If $\rho_{xy.w}$ is close to 0, the linear function is so flat that it will not intersect the function on the right-hand side. That means $\partial\overline{\beta}(r)/\partial r = 0$ has no roots. The equation $\partial\overline{\beta}(r)/\partial r = 0$ has only one root if the two functions on the two sides of Eq. (31) are tangent to each other. That means $\partial^2\overline{\beta}/\partial r^2 = 0$. This critical value of $\rho_{xy.w}$ is named as ξ, which can be solved by $\partial\overline{\beta}(r)/\partial r = 0$ and $\partial^2\overline{\beta}(r)/\partial r^2 = 0$. Moreover, ξ is a function of r_u^2 and r_l^2. □

Proof (*Theorem 3*). By the assumptions in model II, the second moments can be written as

$$\sigma_{yy} = \beta^2\sigma_{x^*x^*} + \gamma'\Sigma_{ww}\gamma + 2\beta\Sigma_{x^*w}\gamma + \sigma_{uu}, \tag{33}$$

$$\sigma_{xy} = \beta\sigma_{xx^*} + \Sigma_{xw}\gamma + \sigma_{\varepsilon u}, \tag{34}$$

$$\Sigma_{wy} = \beta\Sigma_{wx^*} + \Sigma_{ww}\gamma, \tag{35}$$

$$\sigma_{xx} = r^2\sigma_{x^*x^*} + \sigma_{\varepsilon\varepsilon}, \tag{36}$$

$$\Sigma_{xw} = r\Sigma_{x^*w}. \tag{37}$$

Eliminating γ from the system, we have:

$$\sigma_{yy.w} = (\beta/r)^2(r^2\sigma_{x^*x^*} - \sigma_{xx} + \sigma_{xx.w}) + \sigma_{uu}, \tag{38}$$

$$\sigma_{xy.w} = (\beta/r)(r^2\sigma_{x^*x^*} - \sigma_{xx} + \sigma_{xx.w}) + \sigma_{\varepsilon u}, \tag{39}$$

$$\sigma_{xx} = r^2\sigma_{x^*x^*} + \sigma_{\varepsilon\varepsilon}. \tag{40}$$

From $\rho_{\varepsilon u}^2 = \sigma_{\varepsilon u}^2/\sigma_{\varepsilon\varepsilon}\sigma_{uu}$, we can get

$$A\beta^2 + B\beta + C = 0, \tag{41}$$

where

$$A = (r^2 - r_l^2)[(r^2 - r_l^2) + \rho_{\varepsilon u}^2(r_u^2 - r^2)],$$
$$B = -2b(r_u^2 - r_l^2)r(r^2 - r_l^2),$$
$$C = b^2(r_u^2 - r_l^2)^2r^2 - \rho_{\varepsilon u}^2bd(r_u^2 - r_l^2)r^2(r_u^2 - r^2).$$

And $\rho_{\varepsilon u}$ has the same sign as $\sigma_{\varepsilon u}$ so that

$$sign(\rho_{\varepsilon u}) = sign\left\{(r_u^2 - r_l^2)b - (r^2 - r_l^2)\frac{\beta}{r}\right\}. \tag{42}$$

The bounds on r are derived from the condition $B^2 - 4AC \geqslant 0$. The upper bound function of β can be written as

$$\overline{\beta}(r, \rho_{\varepsilon u}^2) = \frac{-B + \sqrt{B^2 - 4AC}}{2A}. \tag{43}$$

Moreover, $\overline{\beta}(r) = \overline{\beta}(r, 1)$. The upper bound on β is solved by

$$\max_{r \geqslant m} \overline{\beta}(r). \tag{44}$$

Since we have the specific form of $\overline{\beta}(r)$, the explicit solution of the maximization problem above does exist and has a complicated and less informative form. Here we only discuss how many maxima the function has on its domain. Note $\overline{\beta}(r_u) = r_u b$, $\overline{\beta}(r) \to \infty$ as $r \to r_l$ and $\overline{\beta}(r)$ is continuous on $r \in (r_l, r_u]$. The behavior of the function $\overline{\beta}(r)$ at the two end points implies that a local maximum, if any, coincides with a local minimum. If there were two local maxima, there would be five different values of r satisfying Eq. (41) for the same value of β. This is impossible because Eq. (41) is a quartic function in r. Therefore, $\overline{\beta}(r)$ is either monotone or has a unique local maximum and a unique local minimum. The maximum is characterized by $\partial\overline{\beta}/\partial r = 0$ and $\partial^2\overline{\beta}/\partial r^2 < 0$. The in-between case is defined as the unique local

maximum and the unique local minimum being the same point. This case is characterized by $\partial\overline{\beta}/\partial r = 0$ and $\partial^2\overline{\beta}/\partial r^2 = 0$. The lower bound on β is much simpler to analyze since the lower bound function $\underline{\beta}(r)$ is a monotonic function.

The tightness of the bounds on β is shown by finding possible values of the unobservables r, σ_{uu}, $\sigma_{\varepsilon\varepsilon}$ and $\sigma_{\varepsilon u}$ which lead to a given value β^* between the bounds (including bounds themselves) in Theorem 3. Obviously, these possible values may not be unique for a given β^*. It is enough to show just one possible case. We let $r = \widetilde{r}$ as follows:

$$\widetilde{r} = \begin{cases} r_l & \text{if } m \leqslant r_l, \\ m & \text{if } \rho_{xy.w} \leqslant \xi \text{ and } m > r_l, \\ m & \text{if } \rho_{xy.w} > \xi,\ r_{\max} \geqslant m > r_l,\ \overline{\beta}(m) \geqslant \overline{\beta}(r_{\max}), \\ m & \text{if } \rho_{xy.w} > \xi,\ r_{\max} \geqslant m > r_l,\ \overline{\beta}(m) < \overline{\beta}(r_{\max}),\ \text{and } \beta^* \leqslant r_u b, \\ r_{\max} & \text{if } \rho_{xy.w} > \xi,\ r_{\max} \geqslant m > r_l,\ \overline{\beta}(m) < \overline{\beta}(r_{\max}),\ \text{and } \beta^* > r_u b, \\ m & \text{if } \rho_{xy.w} > \xi, \text{ and } m > r_{\max}. \end{cases} \tag{45}$$

The first two and the last cases in the definition of $\widetilde{r}$ correspond to the same cases of the bounds on β in Theorem 3. The other three cases of $\widetilde{r}$ correspond to the third case in the theorem. The idea is to find the value of r corresponding to the bounds, which is derived when $\rho^2_{\varepsilon u} = 1$. Then a value of β between the bounds corresponds to a value of $\rho^2_{\varepsilon u}$ in $[0, 1]$ with the value of r fixed. From the derivation above, we have $\max(r_l, m) \leqslant \widetilde{r} \leqslant r_u$. We then let

$$\sigma_{uu} = \sigma_{yy.w} - (\beta^*/\widetilde{r})^2(\widetilde{r}^2\sigma_{x^*x^*} - \sigma_{xx} + \sigma_{xx.w}), \tag{46}$$

$$\sigma_{\varepsilon u} = \sigma_{xy.w} - (\beta^*/\widetilde{r})(\widetilde{r}^2\sigma_{x^*x^*} - \sigma_{xx} + \sigma_{xx.w}), \tag{47}$$

$$\sigma_{\varepsilon\varepsilon} = \sigma_{xx} - \widetilde{r}^2\sigma_{x^*x^*}. \tag{48}$$

From the procedure to find the bounds, we must have $\sigma_{\varepsilon\varepsilon} \geqslant 0$, $\sigma_{uu} \geqslant 0$ and $0 \leqslant \rho^2_{\varepsilon u} \leqslant 1$. Therefore, we find possible values of r, σ_{uu}, $\sigma_{\varepsilon\varepsilon}$ and $\sigma_{\varepsilon u}$ which lead to β^*. Thus, the bounds on β are tight.

The tightness of the bounds on r is shown by finding possible values of β, σ_{uu}, $\sigma_{\varepsilon\varepsilon}$ and $\sigma_{\varepsilon u}$ which lead to a given value r^* such that $\max(r_l, m) \leqslant r^* \leqslant r_u$. We let

$$\sigma_{\varepsilon\varepsilon} = \sigma_{xx} - r^{*2}\sigma_{x^*x^*}, \tag{49}$$

where $\sigma_{\varepsilon\varepsilon} \geqslant 0$ because $r^* \leqslant r_u$. We can easily find σ_{uu}, $\sigma_{\varepsilon u}$, and β satisfying Eqs. (38) through (40). For example, we let $\rho_{\varepsilon u} = 1$ and $\beta = \underline{\beta}$,

$$\widetilde{\beta} = \underline{\beta}(r^*, \rho^2_{\varepsilon u}), \tag{50}$$

$$\sigma_{uu} = \sigma_{yy.w} - (\widetilde{\beta}/r^*)^2(r^{*2}\sigma_{x^*x^*} - \sigma_{xx} + \sigma_{xx.w}), \tag{51}$$

$$\sigma_{\varepsilon u} = \sigma_{xy.w} - (\widetilde{\beta}/r^*)(r^{*2}\sigma_{x^*x^*} - \sigma_{xx} + \sigma_{xx.w}). \tag{52}$$

The derivation of the bounds guarantees that $\sigma_{uu} \geqslant 0$ and $\sigma_{\varepsilon u}/\sqrt{\sigma_{uu}\sigma_{\varepsilon\varepsilon}} = 1$. Therefore, the bounds on r are tight. Moreover, this argument holds not only for $\rho_{\varepsilon u} = 1$, but also for $\rho_{\varepsilon u}$ near 1. This is because the lower bound function $\underline{\beta}(r, \rho^2_{\varepsilon u})$ is continuously differentiable in both r and $\rho^2_{\varepsilon u}$. Since $\underline{\beta}(r, 1)$ is monotonic in r, we must have $\underline{\beta}(r, \rho^2_{\varepsilon u})$ monotonic in r for $\rho_{\varepsilon u}$ near 1. Thus, the above argument is also true for $\rho_{\varepsilon u}$ near 1.

In Theorem 1, we have $r_l = 0$, $m \to 0$, $\xi = 0$ and $r_{\max} = \rho_{xy} r_u$. The definition of $\widetilde{r}$ can be simplified as follows:

$$\widetilde{r} = \begin{cases} \rho_{xy} r_u & \text{if } \beta^* > r_u b, \\ m & \text{if } \beta^* \leqslant r_u b, \end{cases} \tag{53}$$

where m is some positive number close to 0. The tightness of the bounds can be shown in the same way as above. □

Proof (*Theorem 4*). From the explicit expression of $\overline{\beta}(r)$ and $\underline{\beta}(r)$, one can show that $\overline{\beta}(r)/r$ and $\underline{\beta}(r)/r$ are both monotonic in r. Therefore, the bounds can be derived directly, given the range of r. These bounds on γ are the same as in Erickson (1993), since model II can be transformed to resemble the model in Erickson (1993) if rx^* is considered as the latent variable and β/r is its coefficient. The tightness of the bounds also follows the existing results. □

Proof (*Lemma 5*). From the first moment condition $P_x = p + rP_{x^*}$ and $p \geqslant 0$, we have $r \leqslant P_x/P_{x^*}$. Similarly, $q \geqslant 0$ leads to $r \leqslant (1 - P_x)/(1 - P_{x^*})$. Therefore, $r \leqslant r_d$. As in the proof of Theorem 1, the existence of β requires $B^2 - 4AC \geqslant 0$ in Eq. (29), which leads to the bounds on r for different values of $\rho_{\varepsilon u}$, as follows:

$$r \in \begin{cases} (0, r_u] & \text{if } \rho^2_{\varepsilon u} \geqslant \rho^2_{xy}, \\ \left[\sqrt{\frac{\rho^2_{xy} - \rho^2_{\varepsilon u}}{1 - \rho^2_{\varepsilon u}}} r_u, r_u\right] & \text{if } \rho^2_{\varepsilon u} < \rho^2_{xy}. \end{cases} \tag{54}$$

We can write the lower bound on r as a function of $\rho^2_{\varepsilon u}$, say, $r_l(\rho^2_{\varepsilon u})$, as follows:

$$r_l(\rho^2_{\varepsilon u}) = \begin{cases} 0 & \text{if } \rho^2_{\varepsilon u} \geqslant \rho^2_{xy}, \\ \sqrt{\frac{\rho^2_{xy} - \rho^2_{\varepsilon u}}{1 - \rho^2_{\varepsilon u}}} r_u & \text{if } \rho^2_{\varepsilon u} < \rho^2_{xy}. \end{cases} \tag{55}$$

Then, the condition $r_l(\rho^2_{\varepsilon u}) \leqslant r_d$ implies a range of $\rho^2_{\varepsilon u}$. Note that $r_l(\rho^2_{\varepsilon u})$ reaches its maximum $\rho_{xy} r_u$ when $\rho^2_{\varepsilon u} = 0$. From the expression of $r_l(\rho^2_{\varepsilon u})$, one can show that the informative bounds (other than -1 and 1) can only be found in the case of $r_d \leqslant \rho_{xy} r_u$. Therefore, if $r_d > \rho_{xy} r_u$, then $\rho_{\varepsilon u} \in [-1.1]$; if $r_d \leqslant \rho_{xy} r_u$, then $r_l(\rho^2_{\varepsilon u}) \leqslant r_d$ implies $\rho^2_{\varepsilon u} \geqslant (1 - (1 - \rho^2_{xy}) r_u^2/(r_u^2 - r_d^2))$. The sign of $\rho_{\varepsilon u}$ is determined by that of $\sigma_{\varepsilon u}$ so that we have $sign(\rho_{\varepsilon u}) = sign(r_u^2 b - \beta r)$. Since we have explicit forms of function $\overline{\beta}(r)$ and

$\underline{\beta}(r)$, a tedious but straightforward calculation shows that $\rho_{\varepsilon u}$ must be nonnegative if $r \leqslant \rho_{xy} r_u$. Thus, the bounds on $\rho_{\varepsilon u}^2$ lead to those on $\rho_{\varepsilon u}$ directly.

To directly prove the tightness of the bounds on $\rho_{\varepsilon u}$, we need to find the values of r, β, σ_{uu}, $\sigma_{\varepsilon\varepsilon}$ and $\sigma_{\varepsilon u}$ which lead to a particular value of $\rho_{\varepsilon u}$ between the bounds (including the bounds themselves). For a given value $\rho_{\varepsilon u}^*$, we let $r = \widetilde{r}$ and $\beta = \widetilde{\beta}$, as follows:

$$\widetilde{r} = \begin{cases} \frac{1}{2}(\rho_{xy} r_u + r_d) & \text{if } r_d > \rho_{xy} r_u, \\ \dfrac{1}{2}\left(\sqrt{\dfrac{\rho_{xy}^2 - \rho_{\varepsilon u}^{*2}}{1 - \rho_{\varepsilon u}^{*2}}} r_u + r_d\right) & \text{if } r_d \leqslant \rho_{xy} r_u \text{ and } \rho_{\varepsilon u}^* < \rho_{xy}, \\ \frac{1}{2} r_d & \text{if } r_d \leqslant \rho_{xy} r_u \text{ and } \rho_{\varepsilon u}^* \geqslant \rho_{xy}, \end{cases} \tag{56}$$

$$\widetilde{\beta} = \begin{cases} \underline{\beta}(\widetilde{r}, \rho_{\varepsilon u}^{*2}) & \text{if } r_d > \rho_{xy} r_u \text{ and } \rho_{\varepsilon u}^* \geqslant 0, \\ \overline{\beta}(\widetilde{r}, \rho_{\varepsilon u}^{*2}) & \text{if } r_d > \rho_{xy} r_u \text{ and } \rho_{\varepsilon u}^* < 0, \\ \underline{\beta}(\widetilde{r}, \rho_{\varepsilon u}^{*2}) & \text{if } r_d \leqslant \rho_{xy} r_u. \end{cases} \tag{57}$$

The last two cases in the definition of $\widetilde{r}$ correspond to the second case of bounds on $\rho_{\varepsilon u}$ in Lemma 5. In that case, the lower bound on $\rho_{\varepsilon u}$, i.e., $\rho_{\varepsilon u}^* \geqslant (1 - (1 - \rho_{xy}^2) r_u^2 / (r_u^2 - r_d^2))^{1/2}$, implies that $\sqrt{(\rho_{xy}^2 - \rho_{\varepsilon u}^{*2})/(1 - \rho_{\varepsilon u}^{*2})} r_u \leqslant r_d$. The key is to choose an $\widetilde{r}$ such that $\overline{\beta}(\widetilde{r}, \rho_{\varepsilon u}^{*2})$ or $\underline{\beta}(\widetilde{r}, \rho_{\varepsilon u}^{*2})$ changes with $\rho_{\varepsilon u}^*$. In fact, $\widetilde{r}$ can take any value in $(\rho_{xy} r_u, r_d)$ in the first case in the definition of $\widetilde{r}$, any value in $(\sqrt{(\rho_{xy}^2 - \rho_{\varepsilon u}^{*2})/(1 - \rho_{\varepsilon u}^{*2})} r_u, r_d)$ in the second case, and value in $(0, r_d)$ in the third case. The values of σ_{uu}, $\sigma_{\varepsilon\varepsilon}$ and $\sigma_{\varepsilon u}$ can be found as follows:

$$\sigma_{\varepsilon\varepsilon} = \sigma_{xx} - \widetilde{r}^2 \sigma_{x^*x^*}, \tag{58}$$

$$\sigma_{uu} = \sigma_{yy} - \widetilde{\beta}^2 \sigma_{x^*x^*}, \tag{59}$$

$$\sigma_{\varepsilon u} = \sigma_{xy} - \widetilde{\beta}\widetilde{r} \sigma_{x^*x^*}. \tag{60}$$

The derivation of the bound function $\underline{\beta}(r, \rho_{\varepsilon u}^2)$ and $\overline{\beta}(r, \rho_{\varepsilon u}^2)$ guarantees the $\sigma_{uu} \geqslant 0$, $\sigma_{\varepsilon\varepsilon} \geqslant 0$ and $\sigma_{\varepsilon u} / \sqrt{\sigma_{uu} \sigma_{\varepsilon\varepsilon}} = \rho_{\varepsilon u}^*$. Thus, the bounds on $\rho_{\varepsilon u}$ are tight. □

Proof (*Corollary 6*). Since the results in Lemma 5 suggest that $\rho_{\varepsilon u}^2$ can equal 1, we can apply Theorem 1 to the dichotomous model with an extra restriction $r \leqslant r_d$. From Theorem 1, we know that $\underline{\beta}(r)$ is an increasing function and $\overline{\beta}(r)$ reaches its unique maximum at $r_{\max} = \rho_{xy} r_u$. Therefore, if $r_d > \rho_{xy} r_u$, the bounds do not change. Otherwise, the upper bound has to be adjusted to $\overline{\beta}(r_d)$.

As shown in Lemma 5, we must have $\rho_{xy} r_u \leqslant r_d$ and $\beta = \sigma_{xy} / r \sigma_{x^*x^*}$ if $\rho_{\varepsilon u} = 0$. Therefore, the bounds on β follow from the bounds on r.

The tightness of the bounds still holds by the relevant proof of Theorem 3. The major difference between Corollary 6 and Theorem 1 is that there is an additional upper bound on r, i.e., $r \leqslant r_d$. To consider the additional restriction $r \leqslant r_d$, we need to redefine $\tilde{r}$ in the proof of Theorem 3 as follows:

$$\tilde{r} = \begin{cases} \rho_{xy} r_u & \text{if } r_d > \rho_{xy} r_u \text{ and } \beta^* > r_u b, \\ r_d & \text{if } r_d \leqslant \rho_{xy} r_u \text{ and } \beta^* > r_u b, \\ m, & \beta^* \leqslant r_u b, \end{cases} \tag{61}$$

where m is some positive number close to 0. The tightness of the bounds can be shown in the same way as in the proof of Theorem 3. □

References

Angrist, J., Krueger, A., 1999. Empirical strategies in labor economics. In: Ashenfelter, O., Card, D. (Eds.), Handbook of Labor Economics, vol. 3A. North-Holland, Amsterdam, pp. 1277–1366.

Bekker, P., Kapteyn, A., Wansbeek, T., 1984. Measurement error and endogeneity in regression: bounds for ML and 2SLS estimates. In: Dijkstra, T.K. (Ed.), Misspecification Analysis. Springer, Berlin, pp. 85–103.

Bekker, P., Kapteyn, A., Wansbeek, T., 1987. Consistent sets of estimates for regressions with correlated and uncorrelated measurement errors in arbitrary subsets of all variables. Econometrica 55, 1223–1230.

Black, D.A., Berger, M.C., Scott, F.A., 2000. Bounding parameter estimates with nonclassical measurement error. Journal of the American Statistical Association 95, 739–748.

Bollinger, C., 1996. Bounding mean regressions when a binary regressor is mismeasured. Journal of Econometrics 73, 387–399.

Bound, J., Brown, C., Duncan, G.J., Rodgers, W.L., 1994. Evidence on the validity of cross-sectional and longitudinal labor market data. Journal of Labor Economics 12, 345–368.

Bound, J., Brown, C., Mathiowetz, N., 2001. Measurement error in survey data. In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 5. North-Holland, Amsterdam, pp. 3705–3843.

Erickson, T., 1989. Proper posteriors from improper priors for an unidentified errors-in-variables model. Econometrica 57, 1299–1316.

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

Gini, C., 1921. Sull'interpolazione di una retta quando i valori della variabile indipendente sono affetti da errori accidentali. Metroeconomica 1, 63–82.

Iwata, S., 1992. Instrumental variables estimation in errors-in-variables models when instruments are correlated with errors. Journal of Econometrics 53, 297–322.

Kane, T.J., Rouse, C.E., Staiger, D., 1999. Estimating the returns to schooling when schooling is misreported. NBER Working Paper 7235.

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

Klepper, S., 1988b. Regressor diagnostics for the classical errors-in-variables model. Journal of Econometrics 37, 225–250.

Klepper, S., Leamer, E., 1984. Consistent sets of estimates for regression with errors in all variables. Econometrica 52, 163–183.

Koopmans, T.C., 1937. Linear Regression Analysis of Economic Time Series. Haarlem, Bohn.

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

Krasker, W.S., Pratt, J.W., 1987. Bounding the effects of proxy variables on instrumental-variables coefficients. Journal of Econometrics 35, 233–252.

Leamer, E., 1982. Sets of posterior means with bounded variance priors. Econometrica 50, 725–763.

Leamer, E., 1987. Errors in variables in linear systems. Econometrica 55, 893–909.

Patefield, W.M., 1981. Multivariate linear relationships: maximum likelihood estimation and regression bounds. Journal of the Royal Statistical Society B 43, 342–352.

Rodgers, W., Brown, C., Duncan, G., 1993. Errors in survey reports of earnings, hours worked, and hourly wages. Journal of the American Statistical Association 88, 1208–1218.