Economics Letters 136 (2015) 59–63

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Covariate measurement and endogeneity

Daniel L. Millimet *

Southern Methodist University, United States
IZA, Germany

HIGHLIGHTS

- Problems of covariate measurement error and endogeneity are jointly analyzed.
- Reducing measurement error may worsen finite sample performance of IV estimators.
- Stronger instruments diminish the sensitivity of the bias to measurement error.

ARTICLE INFO

Article history:
Received 25 February 2015
Received in revised form
20 August 2015
Accepted 25 August 2015
Available online 1 September 2015

JEL classification:
C36
C81

Keywords:
Measurement error
Endogeneity
Two-Stage Least Squares

ABSTRACT

The effects of improving covariate measurement are investigated when the covariate is endogenous even in the absence of measurement error. Reducing measurement error can exacerbate the finite sample bias of Two-Stage Least Squares. An application reveals this is of practical importance.

1. Introduction

Applied researchers often confront the twin problems of covariate measurement error and endogeneity of the covariate even in the absence of measurement error. The typical solution to covariate measurement error is instrumental variable (IV) estimation. The typical solution to endogeneity (of an accurately measured covariate) is also IV estimation. This, then, begs the question: if a researcher must resort to IV estimation even in the absence of measurement error, is there a gain to improving covariate measurement?

This is an important question for applied researchers as significant effort is often devoted to improving covariate measurement even though the covariate is (presumed) endogenous even in the absence of measurement error. For example, when assessing the impact of environmental regulatory stringency on outcomes such as firm productivity, firm location, and trade flows, considerable effort is allocated to measuring stringency (Levinson, 2001; Brunel and Levinson, 2013; Sauter, 2014). However, stringency is likely to be correlated with unobservables influencing such outcomes, rendering Ordinary Least Squares (OLS) biased even in the absence of measurement error (Millimet and Roy, forthcoming). Similarly, in the returns to education literature, effort is often devoted to circumventing measurement error in self-reported schooling through the use of administrative or transcript data (e.g. Battistin et al., 2014). Again, though, schooling is presumed endogenous even in the absence of measurement error due to omitted innate ability. Other examples abound, from constructing national or subnational indices of employment protection legislation or measures of corruption to firm-level measures of capital stock to individual-level attributes such as permanent income or total consumption.

Here, it is shown that the finite sample bias of the Two-Stage Least Squares (TSLS) estimator may be *exacerbated* by improvement in covariate measurement. Moreover, the sensitivity of the bias to the degree of measurement error – in an absolute sense – is greater when the instruments are weak (another problem often

* Correspondence to: Department of Economics, Box 0496, Southern Methodist University, Dallas, TX 75275-0496, United States. Tel.: +1 214 768 3269; fax: +1 214 768 1821.
E-mail address: millimet@smu.edu.

http://dx.doi.org/10.1016/j.econlet.2015.08.023
0165-1765/

confronted in practice). In sum, researchers using small samples should be cognizant of the potential harm, or at least the possible lack of gain, created by improving the accuracy of covariates that are endogenous even in the absence of measurement error. One should avoid interpreting this result as a rationale for ignoring covariate accuracy. Instead, this points to a more productive use of researcher effort: measurement of strong instruments.

2. Bias

2.1. Preliminaries

Consider the following data generating process (DGP)

$$y = x^*\beta + \varepsilon \tag{1}$$

$$x = x^* + \upsilon \tag{2}$$

$$x^* = z\pi + \eta \tag{3}$$

where y is a $N \times 1$ vector of a dependent variable, x^* is (for simplicity) a $N \times 1$ vector of a correctly measured independent variable, x is a $N \times 1$ vector of the observed independent variable, β is a scalar parameter of interest, z is a $N \times L$ matrix of instrumental variables ($L \geq 1$), π is a $L \times 1$ vector of parameters, and ε, υ, and η are $N \times 1$ vectors of mean zero error terms.[1] The covariance matrix of the errors, Σ, is given by

$$\Sigma = \begin{bmatrix} \sigma_\varepsilon^2 & 0 & \sigma_{\varepsilon\eta} \\ & \sigma_\upsilon^2 & 0 \\ & & \sigma_\eta^2 \end{bmatrix}.$$

In addition, assume that υ is classical measurement error such that $\mathrm{E}[x^{*\prime}\upsilon] = \mathrm{E}[z'\upsilon] = 0$.

The model in (1)–(3) can be written more compactly as

$$y = x\beta + \widetilde{\varepsilon} \tag{4}$$

$$x = z\pi + \widetilde{\eta} \tag{5}$$

where $\widetilde{\varepsilon} \equiv \varepsilon - \beta\upsilon$ and $\widetilde{\eta} \equiv \eta + \upsilon$. Because (4) and (5) comprise a typical system of equations, all the well known results from the literature on OLS and TSLS continue to hold. Specifically, from Hahn and Hausman (2002) and Bun and Windmeijer (2011), the bias of the OLS estimator of β from a regression of y on x is approximately

$$\mathrm{E}[\widehat{\beta}_{OLS}] - \beta \approx \frac{\sigma_{\widetilde{\varepsilon}\widetilde{\eta}}}{\sigma_x^2}. \tag{6}$$

Nagar (1959) and Bun and Windmeijer (2011) provide two different approximations of the bias of the TSLS estimator of β using z to instrument for x. These are given by

$$\mathrm{E}[\widehat{\beta}_{TSLS}] - \beta \approx \frac{\sigma_{\widetilde{\varepsilon}\widetilde{\eta}}}{\pi' z' z \pi}(L-2) \tag{7}$$

$$\mathrm{E}[\widehat{\beta}_{TSLS}] - \beta \approx \frac{\sigma_{\widetilde{\varepsilon}\widetilde{\eta}}}{\sigma_\eta^2 + \sigma_\upsilon^2}\left[\frac{L}{\mu + L} - \frac{2\mu^2}{(\mu+L)^3}\right], \tag{8}$$

respectively, where μ is the concentration parameter (Basmann, 1963) given by

$$\mu \equiv \frac{\pi' z' z \pi}{\sigma_{\widetilde{\eta}}^2}.$$

[1] Utilizing the Frisch–Waugh–Lovell Theorem, other exogenous covariates can be thought of as having been partialled out. Future research might consider multiple endogenous variables, although it seems unlikely that meaningful conclusions could be obtained under general correlation structures for the structural and measurement errors. Moreover, many empirical applications are assumed to contain at most one endogenous regressor.

Table 1
Simulation details.

$L = 3$	$\sigma_\eta^2 = 1$	$\sigma_{x^*}^2 = \sigma_x^2 - \sigma_\upsilon^2$
$N = 100$	$\sigma_\upsilon^2 = \frac{(1-\varphi)\left(\frac{\mu}{N}+1\right)}{1-(1-\varphi)\left(\frac{\mu}{N}+1\right)}\sigma_\eta^2$	$\sigma_\varepsilon^2 = \beta^2\sigma_{x^*}^2$
$\beta = 0.1$	$\sigma_x^2 = \sigma_{\widetilde{\eta}}^2\left(\frac{\mu}{N}+1\right)$	$\sigma_{\varepsilon\eta} = \rho_{\varepsilon\eta}\sigma_\varepsilon\sigma_\eta$

The Nagar approximation requires $\mu \to \infty$ as $N \to \infty$, while the Bun and Windmeijer approximation requires that $\max\{\mu, L\} \to \infty$ as $N \to \infty$.

Utilizing the following approximations

$$\sigma_x^2 \approx \sigma_{\widetilde{\eta}}^2\left(\frac{\mu}{N}+1\right)$$

$$\varphi \equiv 1 - \frac{\sigma_\upsilon^2}{\sigma_x^2} \approx 1 - \frac{\sigma_\upsilon^2}{\sigma_{\widetilde{\eta}}^2\left(\frac{\mu}{N}+1\right)}$$

where φ is the reliability ratio, we can rewrite the three bias approximations in terms of the reliability ratio and the concentration parameter as

$$\mathrm{Bias}_{OLS} \approx \beta(\varphi - 1) + \frac{\sigma_{\varepsilon\eta}}{\sigma_\eta^2 \Gamma_0}\frac{1}{\frac{\mu}{N}+1} \tag{9}$$

$$\mathrm{Bias}_{Nagar} \approx \beta(\varphi-1)\left(\frac{\mu}{N}+1\right)\Gamma_1 + \frac{\sigma_{\varepsilon\eta}}{\sigma_\eta^2\Gamma_0}\Gamma_1 \tag{10}$$

$$\mathrm{Bias}_{BW} \approx \beta(\varphi-1)\left(\frac{\mu}{N}+1\right)\Gamma_2 + \frac{\sigma_{\varepsilon\eta}}{\sigma_\eta^2\Gamma_0}\Gamma_2 \tag{11}$$

where

$$\Gamma_0 \equiv 1 + \frac{(1-\varphi)\left(\frac{\mu}{N}+1\right)}{1-(1-\varphi)\left(\frac{\mu}{N}+1\right)}$$

$$\Gamma_1 \equiv \frac{L-2}{\mu}$$

$$\Gamma_2 \equiv \frac{L}{\mu+L} - \frac{2\mu^2}{(\mu+L)^3}.$$

Note, each bias expression in (9)–(11) contains two terms. The first term in each vanishes in the absence of measurement error ($\varphi \to 1$). The second term in each converges to the usual finite sample bias of OLS or TSLS when a correctly measured covariate is endogenous.

2.2. Covariate accuracy

Of interest here is the effect of reducing the (classical) measurement error in x on the properties of OLS and TSLS when $\sigma_{\varepsilon\eta} \neq 0$. Altering the reliability ratio impacts both terms in each bias expression in (9)–(11). Moreover, a change in the reliability ratio need not impact the two terms in the same direction.

To illustrate the change in bias, Figs. 1–4 plot the Nagar bias and Bun–Windmeijer bias (in absolute value) from (10) and (11), respectively, for selected parameter values. The simulation details are given in Table 1.

L is set to three such that the expectation exists. The variance of ε is chosen such that the population R^2 in (1) is 0.5. The correlation coefficient between ε and η, $\rho_{\varepsilon\eta}$, reflects the degree of endogeneity of x^*. Across Figs. 1–4, $\rho_{\varepsilon\eta}$ varies from 0.5, 0.1, −0.1, and −0.5, respectively. Within each figure, the reliability ratio, φ, is varied from 0.2 to one. In addition, five different values of instrument strength are utilized: $\mu/L \in \{0.1, 0.33, 1, 2, 5\}$. The ratio, μ/L, is the population analog of the first-stage F-statistic (Bound et al., 1995; Stock et al., 2002).

Four salient points are illustrated. First, as seen in Figs. 1 and 2, the bias may be zero in the presence of measurement error. This