Identification of Errors-in-Variables Models Using Repeated Measurements When All the Variables Are Mismeasured

Ping Deng*
Nanchang Institute of Technology
Nanchang, Jiangxi, China

Yingyao Hu†
University of Texas at Austin

November 5, 2006

Abstract

This note uses repeated measurements to identify a latent model of interest with discrete variables subject to measurement errors. We extend the identification results in Li (2002) and Schennach (2004), where the measurement errors are classical. We allow all the variables in the model to be reported with nonclassical errors. The key restriction on the misreporting error may be that people always underreport the true values. The identification does not rely on any parametric specification of the latent model.

JEL classification: C01, C14.
Keywords: measurement error, identification.

1 Introduction

The problem of measurement errors in the survey samples has been studied in economics for a long time. The error is called classical if it is independent of the true values, otherwise, is called nonclassical. Additional sample information used in the identification of continuous errors-in-variables models includes using instruments (see e.g., Wang and Hsiao (1995), Schennach (2006)), repeated measurements (see e.g., Li (2002), Schennach (2004)) or validation data (Chen, X., H. Hong, and E. Tamer (2005)). In the discrete case, Mahajan (2006), Lewbel (2006) and Hu (2005) show that a model may be identified with instruments when

*Department of Economics, Nanchang Institute of Technology, 289 Tianxiang Avenue, Nanchang HI-Tech Development Zone, Nanchang, Jiangxi 330099, China.

†Corresponding author, Department of Economics, University of Texas at Austin, 1 University Station C3100, Austin, TX 78712, 512-475-8556, hu@eco.utexas.edu.

an independent variable is subject to misclassification errors and the dependent variable is error-free.

This note uses repeated measurements to identify a latent model of interest when all the variables are subject to measurement errors. For example, suppose that we are interested in the impact of education on employment. Education is a discrete variable with such possible values as high school, college, and graduate, while employment may take such values as unemployed, part-time, or full-time. Researchers usually use education and employment observed in a survey sample. However, validation studies (see Bound, Brown, and Mathiowetz, 2001) show that education is usually reported with error. The employment variable may also contain reporting errors because people may not want to report exactly how long they work every week.

This note shows that a latent model may be identifiable under suitable assumptions even if all the variables in the model are subject to measurement error. The problem discussed here can also be described as how to identify the distribution of the latent true values if we only observe the joint distribution of two measurements. Li (2002) and Schennach (2004) show that such identification is possible in the classical error case. We show the identification in the case where people always under-report the true values.

The remainder of this note is organized as follows. Section 2 shows where the major problem is. Section 3 presents the main identification results. Section 4 concludes the note.

2 The main problem

Suppose that we are interested in a model containing a variable Z^*, which may be a scalar or a vector. Moreover, we assume that the model of interest is identified if the probability density of Z^*, i.e., f_{Z^*} is identified. In a random sample $\{Z, Z'\}$, we observe repeated measurements Z, Z' of the latent variable Z^*. We assume

Condition 1 $f_{Z|Z^*,Z'} = f_{Z|Z^*}$.

This condition implies that the two measurements are independent conditional on the true values. Condition 1 has been adopted in many relevant studies. We therefore have

$$\begin{aligned} f_{Z',Z}(z',z) &= \int_{\mathcal{Z}} f_{Z|Z^*,Z'}(z|z^*,z')\, f_{Z'|Z^*}(z'|z^*)\, f_{Z^*}(z^*)\, dz^* \qquad (1) \\ &= \int_{\mathcal{Z}} f_{Z|Z^*}(z|z^*)\, f_{Z'|Z^*}(z'|z^*)\, f_{Z^*}(z^*)\, dz^*, \end{aligned}$$

where $\mathcal{Z}$ is the support of Z^*. This equation implies a relationship between the observed density $f_{Z,Z'}$ and the unobserved densities $f_{Z|Z^*}$, $f_{Z'|Z^*}$, and f_{Z^*}. The latent density of interest f_{Z^*} is not identifiable without further restrictions. Condition 1 generalizes the corresponding restrictions on the repeated measurements in Li (2002) and Schennach (2004), where the measurement errors $Z - Z^*$ and $Z' - Z^*$ are independent of each other and the latent true values. To be specific, under the assumption that $Z = Z^* + \epsilon$ and $Z' = Z^* + \epsilon'$ with Z^*, ϵ and ϵ' being mutually independent and having zero mean, one may identify f_{Z^*} from $f_{Z,Z'}$ through the identity

$$\phi_{Z^*}(t) = \exp\left(\int_0^t \frac{\partial \phi_{Z,Z'}(0,t_2)/\partial t_1}{\phi_{Z,Z'}(0,t_2)} dt_2\right),$$

where ϕ stands for the characteristic function of its subscript. Although Li (2002) and Schennach (2004) assume the dependent variable has no measurement errors, their identification results using the identity above can be directly generalized to the case where all the variables are mismeasured.

However, the identification results in the classical error case can not be directly generalized to the nonclassical error case. In other words, the latent model f_{Z^*} is not identified in equation 1 under condition 1. The intuition here is that there are more unknowns on the right-hand side of equation 1 than the observables on the left-hand side. That also means that it is possible to fully identify f_{Z^*} if we impose some restrictions on the error distributions $f_{Z|Z^*}$ and $f_{Z'|Z^*}$ to reduce the number of unknowns. This note shows that the latent density f_{Z^*} is identifiable when people always underreport the true values.

3 Identification with always-under-reporting errors

We first introduce some notations. Without loss of generality, we assume the support of Z and Z' is $\mathcal{Z} = \{\delta_1, \delta_2, \delta_3\}$ with $\delta_1 > \delta_2 > \delta_3$. The ressult here can be easily generalized to the the case where $\mathcal{Z}$ contains a finite number of possible values. We define

$$\mathbf{F}_{Z'|Z^*} = \begin{pmatrix} f_{Z'|Z^*}(\delta_1|\delta_1) & f_{Z'|Z^*}(\delta_2|\delta_1) & f_{Z'|Z^*}(\delta_3|\delta_1) \\ f_{Z'|Z^*}(\delta_1|\delta_2) & f_{Z'|Z^*}(\delta_2|\delta_2) & f_{Z'|Z^*}(\delta_3|\delta_2) \\ f_{Z'|Z^*}(\delta_1|\delta_3) & f_{Z'|Z^*}(\delta_2|\delta_3) & f_{Z'|Z^*}(\delta_3|\delta_3) \end{pmatrix},$$

$$\mathbf{F}_{Z|Z^*} = \begin{pmatrix} f_{Z|Z^*}(\delta_1|\delta_1) & f_{Z|Z^*}(\delta_2|\delta_1) & f_{Z|Z^*}(\delta_3|\delta_1) \\ f_{Z|Z^*}(\delta_1|\delta_2) & f_{Z|Z^*}(\delta_2|\delta_2) & f_{Z|Z^*}(\delta_3|\delta_2) \\ f_{Z|Z^*}(\delta_1|\delta_3) & f_{Z|Z^*}(\delta_2|\delta_3) & f_{Z|Z^*}(\delta_3|\delta_3) \end{pmatrix},$$

$$\mathbf{F}_{Z^*} = \begin{pmatrix} f_{Z^*}(\delta_1) & 0 & 0 \\ 0 & f_{Z^*}(\delta_2) & 0 \\ 0 & 0 & f_{Z^*}(\delta_3) \end{pmatrix},$$

$$\mathbf{F}_{Z',Z} = \begin{pmatrix} f_{Z',Z}(\delta_1,\delta_1) & f_{Z',Z}(\delta_2,\delta_1) & f_{Z',Z}(\delta_3,\delta_1) \\ f_{Z',Z}(\delta_1,\delta_2) & f_{Z',Z}(\delta_2,\delta_2) & f_{Z',Z}(\delta_3,\delta_2) \\ f_{Z',Z}(\delta_1,\delta_3) & f_{Z',Z}(\delta_2,\delta_3) & f_{Z',Z}(\delta_3,\delta_3) \end{pmatrix}.$$

Notice that the matrices contain the same information as their corresponding densities. It is straightforward to show that equation 1 is equivalent to

$$\mathbf{F}_{Z',Z} = \mathbf{F}_{Z|Z^*}^T \times \mathbf{F}_{Z^*} \times \mathbf{F}_{Z'|Z^*}. \tag{2}$$

In this note, we consider the case where people always under-report the true values. This case is particularly useful when we study the reporting error in self-reported information on sensitive issues, such as drug use and smoking behavior, because people tend to report the truth if they do not use drug or do not smoke. Given $\delta_1 > \delta_2 > \delta_3$, we assume

Condition 2 $f_{Z|Z^*}(\delta_i|\delta_j) = f_{Z'|Z^*}(\delta_i|\delta_j) = 0$ *if* $\delta_i > \delta_j$.

This condition implies that

$$\mathbf{F}_{Z'|Z^*} = \begin{pmatrix} f_{Z'|Z^*}(\delta_1|\delta_1) & f_{Z'|Z^*}(\delta_2|\delta_1) & f_{Z'|Z^*}(\delta_3|\delta_1) \\ 0 & f_{Z'|Z^*}(\delta_2|\delta_2) & f_{Z'|Z^*}(\delta_3|\delta_2) \\ 0 & 0 & f_{Z'|Z^*}(\delta_3|\delta_3) \end{pmatrix},$$

$$\mathbf{F}_{Z|Z^*} = \begin{pmatrix} f_{Z|Z^*}(\delta_1|\delta_1) & f_{Z|Z^*}(\delta_2|\delta_1) & f_{Z|Z^*}(\delta_3|\delta_1) \\ 0 & f_{Z|Z^*}(\delta_2|\delta_2) & f_{Z|Z^*}(\delta_3|\delta_2) \\ 0 & 0 & f_{Z|Z^*}(\delta_3|\delta_3) \end{pmatrix}.$$

Moreover, we assume that all the matrices are invertible:

Condition 3 $\mathbf{F}_{Z',Z}$ *has the full rank.*

This assumption implies that all the matrices in equation 2 have the full rank. To be specific, assumption 3 implies that people are willing to report the true values, i.e., $f_{Z|Z^*}(\delta_i|\delta_i) > 0$, $f_{Z'|Z^*}(\delta_i|\delta_i) > 0$ and $f_{Z^*}(\delta_i) > 0$ for all δ_i. Notice that this condition can be tested using observed data because we observe Z and Z' in the sample. We define

$$\mathbf{D}_{Z|Z^*} = \begin{pmatrix} f_{Z|Z^*}(\delta_1|\delta_1) & 0 & 0 \\ 0 & f_{Z|Z^*}(\delta_2|\delta_2) & 0 \\ 0 & 0 & f_{Z|Z^*}(\delta_3|\delta_3) \end{pmatrix},$$

which is a diagonal matrix containing the diagonal elements of the matrix $\mathbf{F}_{Z|Z^*}$. We then have

$$\begin{aligned} \mathbf{F}_{Z',Z} &= \left(\mathbf{D}_{Z|Z^*}^{-1} \times \mathbf{F}_{Z|Z^*}\right)^T \times \left(\mathbf{D}_{Z|Z^*} \times \mathbf{F}_{Z^*} \times \mathbf{F}_{Z'|Z^*}\right) \\ &\equiv L \times U \end{aligned} \tag{3}$$

where $L = \left(\mathbf{D}_{Z|Z^*}^{-1} \times \mathbf{F}_{Z|Z^*}\right)^T$ and $U = \mathbf{D}_{Z|Z^*} \times \mathbf{F}_{Z^*} \times \mathbf{F}_{Z'|Z^*}$. To be specific, we have

$$L = \begin{pmatrix} 1 & \frac{f_{Z|Z^*}(\delta_2|\delta_1)}{f_{Z|Z^*}(\delta_1|\delta_1)} & \frac{f_{Z|Z^*}(\delta_3|\delta_1)}{f_{Z|Z^*}(\delta_1|\delta_1)} \\ 0 & 1 & \frac{f_{Z|Z^*}(\delta_3|\delta_2)}{f_{Z|Z^*}(\delta_2|\delta_2)} \\ 0 & 0 & 1 \end{pmatrix}^T$$

and

$$U = \begin{pmatrix} f_{Z|Z^*}(\delta_1|\delta_1)f_{Z^*}(\delta_1) & 0 & 0 \\ 0 & f_{Z|Z^*}(\delta_2|\delta_2)f_{Z^*}(\delta_2) & 0 \\ 0 & 0 & f_{Z|Z^*}(\delta_3|\delta_3)f_{Z^*}(\delta_3) \end{pmatrix} \times$$
$$\times \begin{pmatrix} f_{Z'|Z^*}(\delta_1|\delta_1) & f_{Z'|Z^*}(\delta_2|\delta_1) & f_{Z'|Z^*}(\delta_3|\delta_1) \\ 0 & f_{Z'|Z^*}(\delta_2|\delta_2) & f_{Z'|Z^*}(\delta_3|\delta_2) \\ 0 & 0 & f_{Z'|Z^*}(\delta_3|\delta_3) \end{pmatrix}.$$

That means L is a unit lower triangular matrix ("unit" refers to ones on the diagonal) and U is an upper triangular one.

Next, we show that L and U are actually uniquely determined by $\mathbf{F}_{Z',Z}$. We may show the uniqueness of the decomposition, i.e., the so-called LU decomposition, by contradiction. Suppose we have unit lower triangular matrices L_1 and L_2 and upper triangular matrices U_1 and U_2, which are observationally equivalent, i.e., $L_1U_1 = L_2U_2$. We then have $L_2^{-1}L_1 = U_2U_1^{-1}$. It is known that the inverse of a unit lower triangular matrix, say L_2^{-1}, is still a unit lower triangular matrix and that the product of two unit lower triangular matrices, say $L_2^{-1}L_1$, is still a unit lower triangular matrix. Notice that $U_2U_1^{-1}$ is an upper triangular matrix. Then, we must have $L_2^{-1}L_1 = U_2U_1^{-1} = I$, where I is the identity matrix. Therefore, we have $L_1 = L_2$ and $U_1 = U_2$. That means the matrices L and U are uniquely determined in equation 3.

Since $f_{Z|Z^*}$ and $f_{Z'|Z^*}$ are conditional densities, we have $\mathbf{F}_{Z|Z^*} \times \mathbf{i} = \mathbf{i}$ and $\mathbf{F}_{Z'|Z^*} \times \mathbf{i} = \mathbf{i}$ with $\mathbf{i} = (1,1,1)^T$. That means

$$\mathbf{D}_{Z|Z^*} = \left[Diag\left(L^T \times \mathbf{i}\right)\right]^{-1},$$

and

$$\mathbf{D}_{Z|Z^*} \times \mathbf{F}_{Z^*} = Diag\left(U \times \mathbf{i}\right),$$

where the function $Diag(\cdot)$ maps a column vector to a diagonal matrix whose diagonal entries equal corresponding entries of the vector. Therefore, the density of interest f_{Z^*} is

identified through

$$\mathbf{F}_{Z^*} = Diag\left(L^T \times \mathbf{i}\right) \times Diag\left(U \times \mathbf{i}\right).$$

We may also identify $f_{Z|Z^*}$ and $f_{Z'|Z^*}$ from

$$\begin{aligned} \mathbf{F}_{Z|Z^*} &= \left[Diag\left(L^T \times \mathbf{i}\right)\right]^{-1} \times L^T, \\ \mathbf{F}_{Z'|Z^*} &= \left[Diag\left(U \times \mathbf{i}\right)\right]^{-1} \times U. \end{aligned}$$

We summarized the identification results in the following theorem

Theorem 1 *Suppose that conditions 1-3 hold. Then, the density $f_{Z',Z}$ uniquely determines f_{Z^*} together with $f_{Z|Z^*}$ and $f_{Z'|Z^*}$.*

This result implies that the latent density of interest may be identified if we observe two repeated measurements and if people always under-report the true values.

4 Conclusion

This note uses repeated measurements to identify a latent model of interest when all the variables are subject to measurement errors. The results extend the identification results in Li (2002) and Schennach (2004) to a nonclassical measurement error case. The identification does not rely on any parametric assumptions on the latent model. Moreover, the restrictions on the measurement error are reasonable in some applications. Although we only show the discrete case, the idea could lead to interesting results in the continuous case in the future research.

References

[1] Bound, J., C. Brown, and N. Mathiowetz, 2001, "Measurement error in survey data," in Handbook of Econometrics, J. J. Heckman and E. Leamer, eds., vol 5.

[2] Chen, X., H. Hong, and E. Tamer, 2005, "Measurement Error Models with Auxiliary Data," Review of Economic Studies, 72, 343-366.

[3] Hu, Y., 2005, "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables," Working paper, University of Texas at Austin.

[4] Hu, Y., and G. Ridder, 2004, "Estimation of Nonlinear model with mismeasured regressor using marginal information," memo.

[5] Lewbel, A., 2006, "Estimation of average treatment effects with misclassification," Econometrica, forthcoming.

[6] Li, T., 2002, "Robust and consistent estimation of nonlinear errors-in-variables models,"Journal of Econometrics, 110, 1-26.

[7] Mahajan, A. 2006, "Identification and estimation of single index models with misclassified regressors," Econometrica.

[8] Newey, W., 2001, "Flexible simulated moment estimation of nonlinear errors-in-variables models," Review of Economics and Statistics, 83(4), pp. 616-627.

[9] Schennach, S., 2004, "Estimation of nonlinear models with measurement error," Econometrica, vol. 72, no 1, pp. 33-76.

[10] Schennach, S., 2006, "Instrumental variable estimation of nonlinear errors-in-variables models," Econometrica, forthcoming.

[11] Wang, L., and C. Hsiao, 1995, "A simulation-based semiparametric estimation of nonlinear errors-in-variables models," Working paper, University of Southern California.