Econometric Theory, **28**, 2012, 696–703.
doi:10.1017/S0266466611000715

MEASUREMENT ERRORS AND CENSORED STRUCTURAL LATENT VARIABLES MODELS

SONGNIAN CHEN
Hong Kong University of Science and Technology

CHENG HSIAO
University of Southern California
Nanyang Technological University
and
City University of Hong Kong

LIQUN WANG
University of Manitoba

We consider censored structural latent variables models where some exogenous variables are subject to additive measurement errors. We demonstrate that overidentification conditions can be exploited to provide natural instruments for the variables measured with errors, and we propose a two-stage estimation procedure. The first stage involves substituting available instruments in lieu of the variables that are measured with errors and estimating the resulting reduced form parameters using consistent censored regression methods. The second stage obtains structural form parameters using the conventional linear simultaneous equations model estimators.

1. INTRODUCTION

In this paper we consider structural latent variables models where the response variables are censored and explanatory variables are measured with errors. Latent variables models have been widely used in econometrics (e.g., Amemiya, 1973, 1974; Maddala, 1983). Most authors in the literature treat the latent variables models as some sort of reduced form specification. However, many variables are jointly dependent and also censored.

We consider models where the joint dependence or simultaneity is in the latent structure, i.e., the latent variables follow the Cowles Commission structural form (e.g., Koopmans and Hood, 1953). It is the observed values that are censored.

We are deeply appreciative of the very helpful comments of a co-editor and two referees. We also thank T. Amemiya for helpful comments. This research is partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Part of this work was carried out while the first author was at the National University of Singapore. Address correspondence to Cheng Hsiao, Dept. of Economies, University of Southern California, Los Angeles, CA 90089-0253, USA; e-mail: chsiao@usc.edu.

 © Cambridge University Press 2011

https://doi.org/10.1017/S0266466611000715 Published online by Cambridge University Press

Measurement errors further complicate the issues. Linear errors-in-variables models have been extensively studied in the literature (see, e.g., the survey of Aigner, Hsiao, Kapteyn, and Wansbeek, 1984). However, censoring introduces noninvertibility between the latent variables and observables, which leads to nonlinearity in the observed data. Nonlinear errors-in-variables models raise complicated identification and estimation issues (see, e.g., Carroll, Ruppert, Stefanski, and Crainiceanu, 2006; Wang and Hsiao, 2011). In this paper, we explore the specific structures of our models to derive identification conditions and root-n consistent and asymptotically normal estimators. We also follow the literature on linear simultaneous equations models with measurement error (e.g., Geraci, 1976; Hsiao, 1976, 1977, 1979) to explore the trade-off between overidentification of a model with the underidentification due to measurement errors.

Section 2 introduces our model, and Section 3 considers the issues of identification and estimation. Concluding remarks are in Section 4.

2. THE MODEL

Let $(y_i', x_i', z_i'), i = 1, 2, ..., n$, be $m + k + \ell$ observed variables and let (y_i', x_i') be related to latent variables $(y_i^{*'}, x_i^{*'})$ in the form

$$y_g = y_g^* \mathbf{1}(y_g^* > 0), \qquad g = 1, ..., m, \tag{2.1}$$

$$x = x^* + \delta, \tag{2.2}$$

where y_g and y_g^* denote the gth component of y and y^*, respectively, and $\mathbf{1}(A) = 1$ if A occurs and 0 otherwise. We make the following assumptions.

A1. The $k \times 1$ random vectors $\{x_i^*, i = 1, 2, ..., n\}$ are independent and identically distributed with finite sixth-order moments, and $\mathrm{E}(x_i^* x_i^{*'})$ is nonsingular.

A2. The measurement error δ is independent of (x^*, z) and is symmetrically distributed with mean zero and covariance matrix Σ_δ.

A3. The latent variables x^* are related to z through

$$x^* = Az + \varepsilon, \tag{2.3}$$

where A is a $k \times \ell$ constant matrix with $\mathrm{rank}(A) = k$ and ε is independent of (z, δ) and symmetrically distributed with mean zero and covariance matrix Σ_ε.

Remark 2.1. Note that in A2 Σ_δ needs not to be nonsingular. If one or more components of x^* are directly observed, then the corresponding diagonal elements of Σ_δ can be zero. However, for ease of exposition, we assume Σ_δ to be nonsingular. Also note that although the relationship (2.3) follows from the usual instrumental variables assumption on z, it is a relatively strong condition. For instance, if x^* is discrete, then A3 is unlikely to be satisfied.

https://doi.org/10.1017/S0266466611000715 Published online by Cambridge University Press

We assume the data generating process for y^* takes the form[1]

$$y^* = \Gamma y^* + Bx^* + u, \tag{2.4}$$

where Γ is an $m \times m$ constant matrix with diagonal elements equal to zero and B is an $m \times k$ constant matrix. We make the following assumptions.

A4. The parameters (Γ, B, A) lie in the interior of a compact space Ω.

A5. The error term u is independent of $(x^*, \delta, \varepsilon)$ and symmetrically distributed with mean zero and nonsingular covariance matrix Σ_u.

Model (2.4) is in the form of the Cowles Commission structural equations model. To ensure that given x^* there is a one-to-one correspondence between a random draw of u from its distribution $f(u)$ and y^* in model (2.4), we make the following assumption.

A6. The determinant $\det(I_m - \Gamma) \neq 0$, where I_m is the $m \times m$ identity matrix.

3. IDENTIFICATION AND ESTIMATION

Let the structure of model (2.4) be denoted by $F = (I - \Gamma, -B)$ where f_g' denotes the gth row of F. We assume that f_g' is subject to the linear restrictions in the form

$$f_g' \Phi_g = 0, \tag{3.1}$$

where Φ_g is an $(m+k) \times r_g$ known matrix with full column rank. Then a standard necessary and sufficient condition for the identification of the gth equation in (2.4) from $(y^{*'}, x^{*'})$ is as follows.

A7. $\operatorname{rank}(F\Phi_g) = m - 1$.

If all equations in (2.4) satisfy A7, the whole model (2.4) is identified. If we observe x in (2.2) rather than x^*, A7 may not be sufficient to identify (2.4). However, identifiability can be achieved if instrumental variables are available. Because y^* is censored, there is no one-to-one correspondence between u and y; we derive the identifiability from the existence of consistent estimators.

PROPOSITION 3.1. *Under A1, A2, and A4–A6, the sufficient conditions for the identification of the gth equation of (2.4) are A3 and A7.*

Proof. Substituting (2.3) into (2.4) yields

$$\begin{aligned} y^* &= (I - \Gamma)^{-1} BAz + (I - \Gamma)^{-1}(u + B\varepsilon) \\ &= \Pi z + \upsilon, \end{aligned} \tag{3.2}$$

where $\Pi = \Pi^* A$ and $\Pi^* = (I - \Gamma)^{-1} B$. Because u and ε are symmetric about zero, $-\upsilon = (I - \Gamma)^{-1}(-u + B(-\varepsilon))$ has the same distribution as υ, which implies that υ is also symmetric about zero. Therefore standard single equation

https://doi.org/10.1017/S0266466611000715 Published online by Cambridge University Press

censored estimation methods (e.g., Amemiya, 1973, 1974; Powell, 1984, 1986) can be used equation by equation to obtain a consistent estimator of Π. Furthermore, by A2 and A3 δ is independent of z, and hence A can be identified by $A = \mathrm{E}(xz')[\mathrm{E}(zz')]^{-1}$. It follows that the consistent estimator of Π^* can be solved as

$$\Pi^* = \Pi A'(AA')^{-1} \tag{3.3}$$

because AA' is nonsingular by A3. Given that $\Pi^* = (I - \Gamma)^{-1}B$ is known, the identification conditions follow straightforwardly from standard derivation. □

Under the additional assumption that the random measurement errors are uncorrelated (i.e., Σ_δ is diagonal), the gth equation may be identified without the presence of additional instruments (i.e., $\ell = 0$) because the excluded exogenous variables can be used as instruments. Therefore, instead of A2 we have the following assumption.

A2′. The measurement error δ is independent of (x^*, z) and is symmetrically distributed with mean zero and diagonal covariance matrix Σ_δ.

Then equations (2.2) and (2.4) imply the reduced form

$$\begin{aligned} y^* &= \Pi^* x + (I - \Gamma)^{-1}u - \Pi^*\delta \\ &= \Pi^*\Sigma_\delta\Sigma_x^{-1}\mu_x + \Pi^*(I - \Sigma_\delta\Sigma_x^{-1})x + \tilde{v} \\ &= \tilde{\pi} + \tilde{\Pi}x + \tilde{v}, \end{aligned} \tag{3.4}$$

where $\mu_x = \mathrm{E}(x)$ and $\tilde{v} = (I - \Gamma)^{-1}u - \Pi^*(\delta - \Sigma_\delta\Sigma_x^{-1}(x - \mu_x))$. It is easy to verify that $\tilde{v}$ has zero mean and is uncorrelated with x. For each $1 \le g \le m$, the reduced form equation for y_g^* is given by

$$y_g^* = \tilde{\pi}_g + \tilde{\Pi}_g x + \tilde{v}_g \tag{3.5}$$

independent of whether other elements of y^* are zero or positive. In other words, censoring on y_g^* happens only if $\tilde{v}_g < -\tilde{\pi}_g - \tilde{\Pi}_g x$. Therefore we can ignore the correlations across equations and try to obtain a consistent estimator of $\tilde{\Pi} = \Pi^*(I - \Sigma_\delta\Sigma_x^{-1})$ equation by equation. Given $\tilde{\Pi}$, we can derive the identification conditions similarly as in Hsiao (1976) if the included elements of x^* are correlated with the excluded elements of x^*.

To implement the estimators of Amemiya (1973) or Powell (1984, 1986), we impose an additional condition.

A3′. The conditional distribution of $\tilde{v}$ given x is symmetric about 0, and the elements of x^* that are included in the first equation are correlated with the other elements of x^*.

Note that by construction the conditional distribution $f(\tilde{v}|x)$ is symmetric when (x^*, δ, u) have an elliptical distribution. The family of elliptical distributions contains many commonly seen distributions such as uniform, normal,

https://doi.org/10.1017/S0266466611000715 Published online by Cambridge University Press

Cauchy, Student's t, double exponential, and many other distributions. See, e.g., Fang and Zhang (1990).

To illustrate how excluded exogenous variables can be used as instruments for the included exogenous variables that are measured with error, without loss of generality we consider the case where $g = 1$ and the prior restrictions are in the form of exclusion restrictions (e.g., Fisher, 1966; Hsiao, 1983). For notational ease, we assume that all m endogenous variables appear in the first equation and the exclusion restrictions are in the form

$$f_1' = (\tilde{\gamma}_1', \beta_1', 0_{k_2}'), \tag{3.6}$$

where $\tilde{\gamma}_1 = (1, \gamma_1')'$ is $m \times 1$, β_1 is $k_1 \times 1$, 0_{k_2} is a $k_2 \times 1$ vector of zeros, and $k_1 + k_2 = k$. Correspondingly, we partition the vectors as $x^* = (x_1^{*'}, x_2^{*'})'$, $x = (x_1', x_2')'$ and the covariance matrix as

$$\Sigma_x = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then we can write the linear projection of x_1^* onto x_2 as

$$\begin{aligned} x_1^* &= \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 + \Sigma_{12}\Sigma_{22}^{-1}x_2 + \varepsilon_1 \\ &= a_1 + A_1 x_2 + \varepsilon_1, \end{aligned} \tag{3.7}$$

where $\mu_j = \mathrm{E}(x_j)$, $j = 1, 2$. Further, from A2′ it is easy to verify by construction that ε_1 has zero mean and is uncorrelated with x_2. Substituting (3.7) into the reduced form of (2.4) yields

$$\begin{aligned} y^* &= (\Pi_1^*, \Pi_2^*)\begin{pmatrix} x_1^* \\ x_2^* \end{pmatrix} + (I - \Gamma)^{-1}u \\ &= \tilde{\pi}_2 + \tilde{\Pi}_2^*(x_2 - \mu_2) + \tilde{v}_2, \end{aligned} \tag{3.8}$$

where $\tilde{\pi}_2 = \Pi_1^*\mu_1 + \Pi_2^*\mu_2$, $\tilde{\Pi}_2^* = \Pi_1^* A_1 + \Pi_2^*(I - \Sigma_{\delta_2}\Sigma_{22}^{-1})$, and

$$\tilde{v}_2 = (I - \Gamma)^{-1}u + \Pi_1^*\varepsilon_1 - \Pi_2^*\delta_2 + \Pi_2^*\Sigma_{\delta_2}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

is symmetrically distributed conditional on x_2 under A3′. Further, the prior restrictions on the first equation

$$\tilde{\gamma}_1'\Pi_1^* = \beta_1', \qquad \tilde{\gamma}_1'\Pi_2^* = 0 \tag{3.9}$$

imply

$$\tilde{\gamma}_1'\tilde{\Pi}_2^* = \beta_1' A_1. \tag{3.10}$$

On the other hand, the reduced form of y^* conditional on x takes the form of (3.4) with

$$\tilde{\Pi} = \Pi^*(I_k - \Sigma_\delta \Sigma_x^{-1}) = (\tilde{\Pi}_1, \tilde{\Pi}_2), \tag{3.11}$$

$$\tilde{\Pi}_1 = \Pi_1^*(I_{k_1} - \Sigma_{\delta_1}\Sigma^{11}) - \Pi_2^*\Sigma_{\delta_2}\Sigma^{21}, \tag{3.12}$$

$$\tilde{\Pi}_2 = \Pi_2^*(I_{k_2} - \Sigma_{\delta_2}\Sigma^{22}) - \Pi_1^*\Sigma_{\delta_1}\Sigma^{12}, \tag{3.13}$$

https://doi.org/10.1017/S0266466611000715 Published online by Cambridge University Press

where Σ^{11}, Σ^{12}, Σ^{21}, and Σ^{22} are the corresponding components of the inverse matrix of Σ_x, i.e.,

$$\Sigma_x^{-1} = \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix}.$$

Let

$$\begin{aligned} H &= \tilde{\Pi}_2^* - \tilde{\Pi}_1 A_1 + \tilde{\Pi}_2 \Sigma^{21} (\Sigma^{12} \Sigma^{21})^{-1} \Sigma^{11} A_1 \\ &= \Pi_2^* \Big[I - \Sigma_{\delta_2} \Sigma_{22}^{-1} + \Sigma_{\delta_2} \Sigma^{21} A_1 + (I - \Sigma_{\delta_2} \Sigma^{22}) \Sigma^{21} \\ &\qquad \times (\Sigma^{12} \Sigma^{21})^{-1} \Sigma^{11} A_1 \Big]. \end{aligned} \tag{3.14}$$

Then the prior restriction $\tilde{\gamma}_1' \Pi_2^* = 0$ implies

$$\tilde{\gamma}_1' H = 0. \tag{3.15}$$

Further, note that the matrices Σ_x and Σ_x^{-1} can be consistently estimated using the observations on x. Thus, the model parameters can be consistently estimated using the following procedure.

Step 1. Applying the single equation censored regression methods of Amemiya (1973) or Powell (1984, 1986) to each equation of (3.4) to obtain a consistent estimator $\hat{\tilde{\Pi}}$ of $\tilde{\Pi} = \Pi^*(I_k - \Sigma_\delta \Sigma_x^{-1}) = (\tilde{\Pi}_1, \tilde{\Pi}_2)$.

Step 2. Applying censored regression methods to (3.8) yields $\hat{\tilde{\Pi}}_2^*$, which converges to $\tilde{\Pi}_2^* = \Pi_1^* A_1 + \Pi_2^* (I - \Sigma_{\delta_2} \Sigma_{22}^{-1})$.

Step 3. Regressing x_1 on x_2 yields $\hat{A}_1$, which converges to A_1.

Step 4. Substituting the preceding estimators into (3.14) to obtain a consistent estimator $\hat{H}$ of H; then obtaining the consistent estimators of $\tilde{\gamma}_1$ and β_1 through a minimum distance estimation procedure based on the prior restrictions (3.15) and (3.10).

Note that in step 4 applying the least squares method yields the censored analogue of the two-stage least squares of a linear simultaneous equations model. Therefore we have the following results.

PROPOSITION 3.2. *Under A1, A2′, A3′, and A4–A7, the first equation of (2.4) is identifiable if the number of overidentifying restrictions* $k_2 \geq k_1 + m - 1$.

Because both the excluded variables and z can serve as instruments for the first equation ($g = 1$), it follows from the preceding result and those of Hsiao (1976) that when $k_2 < k_1 + m - 1$, additional instruments are needed to identify the first equation.

PROPOSITION 3.3. *Under the conditions of Proposition 3.2, a necessary condition to identify the first equation is that there exist at least* $m - 1 + k_1 - k_2$ *additional instruments.*

https://doi.org/10.1017/S0266466611000715 Published online by Cambridge University Press

In general if only $m_1 < m$ endogenous variables appear in the first equation so that $\tilde{\gamma}_1$ in (3.6) contains $m - m_1$ zeros, then in Propositions 3.2 and 3.3 m should be replaced by m_1.

4. CONCLUSION

We discussed the identification and estimation of censored latent structural models with measurement errors. We postulate a latent structure in the spirit of the Cowles Commission structural equation model. It is the observed values that are censored. We showed that identification and consistent estimation can be achieved using instrumental variables. We also showed that, as in the linear simultaneous equations model with measurement errors, the overidentification conditions could be used to trade off the unidentification due to measurement errors. However, because of the censoring efficient estimation of the structural model becomes very complicated. In this paper we proposed a simple and easy to implement multistep consistent estimator.

NOTE

1. Amemiya (1979) considered a mixture case where y_1^* and y_2^* depended on each other as in (2.4); however, the observed $y_1 = y_1^*$ and $y_2 = y_2^* \mathbf{1}(y_2^* > 0)$. Because the reduced forms for y_1^* and y_2^* are identical to the reduced form of (2.4), our two-step procedure can also be applied to this case except that the reduced form for the y_1 equation will be estimated by the least squares method.

REFERENCES

Aigner, D.J., C. Hsiao, A. Kapteyn, & T. Wansbeek (1984) Latent variable models in econometrics. In Z. Griliches & M.D. Intriligator (eds.), *Handbook of Econometrics,* vol. II, pp. 1321–1393. North-Holland.

Amemiya, T. (1973) Regression analysis when the dependent variable is truncated normal. *Econometrica* 41, 997–1016.

Amemiya, T. (1974) Multivariate regression and simultaneous equation models when the dependent variables are truncated normal. *Econometrica* 42, 999–1012.

Amemiya, T. (1979) The estimation of a simultaneous equation Tobit model. *International Economic Review* 20, 169–181.

Carroll, R.J., D. Ruppert, L.A. Stefanski, & C. Crainiceanu (2006) *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. Chapman and Hall.

Fang, K.T. & Y.T. Zhang (1990) *Generalized Multivariate Analysis*. Springer-Verlag.

Fisher, F.M. (1966) *The Identification Problem in Econometrics*. McGraw-Hill.

Geraci, V.J. (1976) Identification of simultaneous equation models with measurement error. *Journal of Econometrics* 4, 263–283.

Hsiao, C. (1976) Identification and estimation of simultaneous equation models with measurement error. *International Economic Review* 17, 319–339.

Hsiao, C. (1977) Identification for a linear dynamic simultaneous error-shock model. *International Economic Review* 18, 181–194.

Hsiao, C. (1979) Measurement error in a dynamic simultaneous equations model with stationary disturbances. *Econometrica* 47, 475–494.

Hsiao, C. (1983) Identification. In Z. Griliches & M.D. Intriligator (eds.), *Handbook of Econometrics,* vol. I, pp. 223–284. North-Holland.

https://doi.org/10.1017/S0266466611000715 Published online by Cambridge University Press

Koopmans, T.C. & W. Hood (1953) The estimation of simultaneous linear economic relationships. In W.C. Hood & T.C. Koopmans (eds.), *Studies in Econometric Methods*, pp. 112–199. Wiley.

Maddala, G.S. (1983) *Limited-Dependent and Qualitative Variables Economics.* Cambridge University Press.

Powell, J.L. (1984) Least absolute deviations estimation of the censored regression models. *Journal of Econometrics* 25, 303–325.

Powell, J.L. (1986) Symmetrically trimmed least squares estimation of Tobit models. *Econometrica* 54, 1435–1460.

Wang, L. & C. Hsiao (2011) Method of moments estimation and identifiability of semiparametric nonlinear errors-in-variables models. *Journal of Econometrics*; doi:10.1016/j.jeconom.2011.05.004.

https://doi.org/10.1017/S0266466611000715 Published online by Cambridge University Press