# *A Consistent Instrumental Variable Estimator for Errors in Covariates in Limited Dependent Variable Models*

Kenneth A. Bollen
Odum Institute
Department of Sociology
University of North Carolina at Chapel Hill

D. Roland Thomas
Sprott School of Business
Carleton University
Canada

Liqun Wang
Department of Statistics
University of Manitoba, Canada

# Outline of Presentation

- Overview & other approaches
- Problem of instrumental variable (IV) approach
- A Consistent IV Estimator
- Asymptotic covariance matrix
- Simulation
- Conclusions

# Overview and approaches

- Ordinal or dichotomous dependent variable
- If covariate available, straightforward probit

$$y^* = \alpha_1 + \alpha_2 \xi_1 + \epsilon \qquad \epsilon \sim N(0,1)$$

$$y = I(y^* \geq 0)$$

# Overview and approaches

- Only proxy covariate ( $x_1$ ) available

$$x_1 = \xi_1 + \delta_1$$

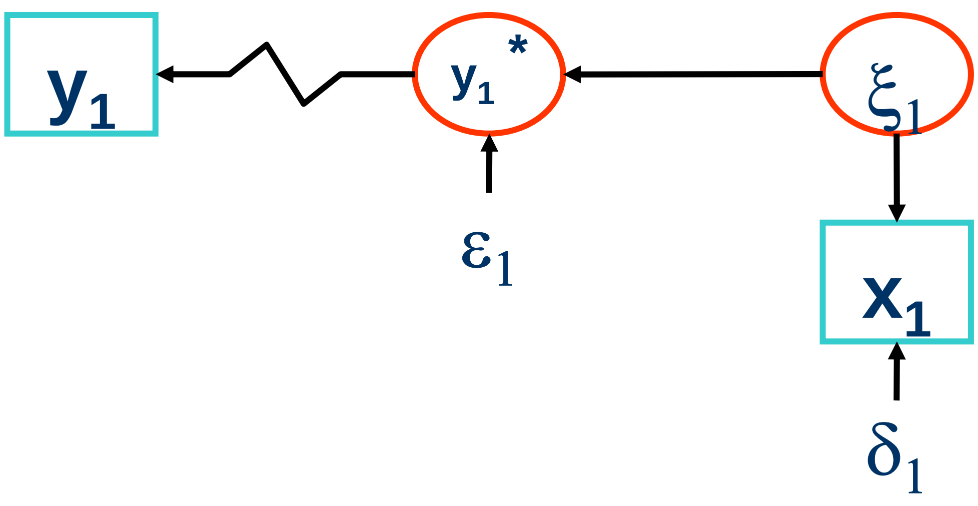
$$\xi_1 = x_1 - \delta_1$$

$$y^* = \alpha_1 + \alpha_2 \xi_1 + \epsilon$$

$$y^* = \alpha_1 + \alpha_2 x_1 - \alpha_2 \delta_1 + \epsilon$$

# Overview and approaches

- Path diagram of equations



# Overview and approaches

- Problems with usual probit regression

$$y^* = \alpha_1 + \alpha_2 x_1 - \alpha_2 \delta_1 + \epsilon$$

$$V(-\alpha_2\delta_1 + \varepsilon) = \alpha_2^2 V(\delta_1) + V(\varepsilon)$$ Not equal to 1

$$C(\delta_1, x_1) \neq 0$$

*plim* $(\widehat{\alpha}_2) \neq \alpha_2$

# Overview and approaches

- Approaches
  - Simulation extrapolation (e.g., Carroll et al., 1996)
  - Regression calibration (e.g., Carroll and Stefanski, 1990; Gleser, 1990)
  - Known reliability of covariate
  - Instrumental variable estimator
  - Above approximately consistent

# Problem of Instrumental Variable (IV) approach

- Suppose we have IV $x_2$ correlated with $\xi_1$ independent of $\delta_1$ and $\epsilon$
- Form $\hat{x}_1 = \beta_1 + \beta_2 x_2$

  and $y^* = \alpha_1 + \alpha_2 \hat{x}_1 + u$ $u = \alpha_2(x_1 - \hat{x}_1) - \alpha_2 \delta_1 + \epsilon$

  - $\hat{x}_1$ and $u$ uncorrelated (solves one problem)

  BUT $V(u) = V(\epsilon) + \alpha_2^2 \left[ \frac{V(\xi_1)V(x_2) - C^2(\xi_1, x_2)}{V(x_2)} \right]$

Not equal to 1, so inconsistent estimator

# A Consistent IV Estimator

- Generalize to multiple covariates, proxies, IVs:

$$y^* = \alpha_1 + \boldsymbol{\alpha}_2'\boldsymbol{\xi} + \varepsilon,$$

$$\mathbf{x} = \boldsymbol{\xi} + \boldsymbol{\delta},$$

$$\mathbf{z} \in R^q$$ IVs

- Usual IV estimator to predict covariates

$$\beta_2 = \Sigma_{\mathbf{zz}}^{-1}\Sigma_{\mathbf{zx}} \qquad \beta_1 = \mu_x - \beta_2'\mu_{\mathbf{z}}$$

# A Consistent IV Estimator

Define

$$\mathbf{u} = \boldsymbol{\xi} - \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2'\mathbf{z}$$

Implies

$$\xi = \beta_1 + \beta_2'\mathbf{z} + \mathbf{u}$$

Substitute into

$$y^* = \alpha_1 + \boldsymbol{\alpha}_2'\boldsymbol{\xi} + \varepsilon,$$

Leading to

$$y^* = \gamma_1 + \boldsymbol{\gamma}_2'\mathbf{z} + v,$$

# A Consistent IV Estimator

$$y^* = \gamma_1 + \boldsymbol{\gamma}_2'\mathbf{z} + v,$$

where $\gamma_1 = \alpha_1 + \boldsymbol{\alpha}_2\boldsymbol{\beta}_1$ $\gamma_2 = \beta_2\boldsymbol{\alpha}_2$ $v = \varepsilon + \boldsymbol{\alpha}_2'\mathbf{u}$

Apply probit to equation at top: $\tilde{\gamma}_1 = \gamma_1/\sigma_v$ $\tilde{\boldsymbol{\gamma}}_2 = \boldsymbol{\gamma}_2/\sigma_v$

$\boldsymbol{\alpha}_2 = \mathbf{M}\boldsymbol{\gamma}_2 = \sigma_v\mathbf{M}\tilde{\boldsymbol{\gamma}}_2$ where $\mathbf{M} = \left(\boldsymbol{\Sigma}_{\mathbf{xz}}\boldsymbol{\Sigma}_{\mathbf{zx}}\right)^{-1}\boldsymbol{\Sigma}_{\mathbf{xz}}\boldsymbol{\Sigma}_{\mathbf{zz}}$

# A Consistent IV Estimator

- We show that

$$\sigma_v^2 = \left(1 + \tilde{\boldsymbol{\gamma}}_2' \boldsymbol{\Sigma}_{zz} \tilde{\boldsymbol{\gamma}}_2 - \left[\sqrt{1 + \tilde{\boldsymbol{\gamma}}_2' \boldsymbol{\Sigma}_{zz} \tilde{\boldsymbol{\gamma}}_2}\,\right] \tilde{\boldsymbol{\gamma}}_2' \mathbf{M}' \boldsymbol{\Sigma}_{xy} / \varphi \right)^{-1}.$$

where $\varphi = \phi\left[\Phi^{-1}(\mu_y)\right]$

$$\alpha_1 = \sigma_v \tilde{\gamma}_1 - \sigma_v \boldsymbol{\beta}_1' \mathbf{M} \tilde{\boldsymbol{\gamma}}_2. \qquad \boldsymbol{\alpha}_2 = \sigma_v \mathbf{M} \tilde{\boldsymbol{\gamma}}_2$$

# Asymptotic covariance matrix

- Use delta method: $$\sqrt{n}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{L} N(0, \sigma_v^2 \mathbf{ABA}')$$

$\mathbf{B}$ usual asymp. cov. matrix of MLE $\tilde{\boldsymbol{\gamma}}$

$$\mathbf{A} = \begin{pmatrix} 1 & -\boldsymbol{\beta}_1'\mathbf{M} + (\tilde{\gamma}_1 - \boldsymbol{\beta}_1'\mathbf{M}\tilde{\boldsymbol{\gamma}}_2)\mathbf{D}/\sigma_v \\ 0 & \mathbf{M} + \mathbf{M}\tilde{\boldsymbol{\gamma}}_2\mathbf{D}/\sigma_v \end{pmatrix}.$$

$$\mathbf{D} = -\frac{1}{2\sigma_v^3}\left(2\tilde{\gamma}_2'\Sigma_{zz} - \left[\sqrt{1+\tilde{\gamma}_2'\Sigma_{zz}\tilde{\gamma}_2}\right]\frac{\Sigma_{xy}'\mathbf{M}}{\varphi} - \frac{\Sigma_{xy}'\mathbf{M}\tilde{\gamma}_2\tilde{\gamma}_2'\Sigma_{zz}}{\varphi\sqrt{1+\tilde{\gamma}_2'\Sigma_{zz}\tilde{\gamma}_2}}\right).$$

# SIMULATION

## Model

$$y^* = \alpha + \beta_1\xi_1 + \beta_2\xi_2 + \beta_3\xi_3 + \zeta$$

$$\xi_i \sim N(0, 1), \quad i = 1, 2, 3, \qquad \zeta \sim N(0, 1)$$

where $\alpha = 0, \beta_1 = \beta_2 = \beta_3 = 1/sqrt(3)$,

and $\zeta$ and $\xi$ 's are independent.

Hence model CD = 0.5

# SIMULATION

Fallible measures of the ξ's

$$x_i = \xi_i + \delta_i , \qquad i = 1, 2, 3.$$

Instrumental variables

$$z_i = \xi_1 + \varepsilon_i , \qquad i = 1, 2$$

$$z_i = \xi_2 + \varepsilon_i , \qquad i = 3, 4$$

$$z_i = \xi_3 + \varepsilon_i , \qquad i = 4, 5$$

$\varepsilon$'s and $\delta$'s mutually independent, $N(0, 1)$, and independent of $\xi$'s and $\zeta$.

# SIMULATION

Bias

$$\overline{B}_i = \overline{\beta}_i - \sqrt{3}/3 \quad \text{where} \quad \overline{\beta}_i = \sum_{j=1}^{1000} \hat{\beta}_{ij} / 1000$$

Mean Bias

$$\overline{B} = \sum_{i=1}^{3} \overline{B}_i / 3$$

MC Standard Error

$$\sigma_{MC}(\overline{B}) = [\sqrt{(1' \hat{\Sigma}_\beta 1)/9}] / \sqrt{1000}$$

# SIMULATION

**3 latent predictors; model CD = 0.5; 3 indicators per latent predictor; CD = 0.5 for each indicator; 2 IV's per latent predictor.**

| Sample Size (N) | $\overline{B}$ | $\sigma_{MC}(\overline{B})$ | Bias Detected? $\lvert\overline{B}\rvert > 2\sigma_{MC}(\overline{B})$ |
|---|---|---|---|
| 100 | 5.3% | 1.5% | Yes |
| 200 | 3.2% | 0.8% | Yes |
| 500 | 1.6% | 0.5% | Yes |
| 1000 | 0.3% | 0.3% | No |

**Mean bias for intercept undetectable even for N = 100**

# Conclusions

- Consistent IV estimator possible
  - No need to settle for "approximately consistent"
- Asymptotic covariance matrix presented
- Simulation illustration shows maximum finite sample bias at small N (=100 , 5%) and minimum at large N (=1000, 0.3%)
- Simulations under broader range of conditions & comparison to other estimators needed