Latent Variable and Measurement Error Modeling in the Social Sciences

D. Roland Thomas
Sprott School of Business
Carleton University

Joint work with

Irene R.R. Lu
School of Administrative Studies
York University

Outline

- The Social Science Approach to LV/EIV
- One Slide Summary of SEM
- Two-Step Methods
  - Predictors / Consistency / Examples
- Analysis of Two Step Bias
- IV/2SLS
  - Bollen’s method / Issues / Example
- Probit Regression with Latent Predictors
  - Consistent Estimates / Simulation

Latent Variables & Measurement Error

Unobservable Variables

- Hypothetical, useful for theory building
- Social / education / business sciences

  ξ = peer relations

  X_1 = I have many friends (1-5)

  X_2 , X_3, X_4, etc…
- Not replicates
- Typically discrete measures

"Latent Variables"

- Real, but difficult to measure (Platonic)
- Physical / medical sciences

  ξ = calorific intake

  X = self-reports on eating
- May have replicate measures
- Often continuous measures

"Errors in Variables"

Both comprise measurement error problems

Measurement Error Models

- Fuller (1987): Measurement Error Models
- Carroll, Rupert, & Stefanski (1995): Measurement Error in Nonlinear Models

Linear: $Y = \beta_0 + \beta_1\xi + \beta_2 Z + \zeta$

Non-linear: $E(Y|\ \xi, Z) = f(\xi, Z, \boldsymbol{\beta})$

$X = \xi + \delta$; $\rho(\zeta, \delta) = 0$; Z exactly measured

- Need side-conditions on unknown variances
- For example, if $\sigma^2(\delta)$ known or estimable, we can estimate $\boldsymbol{\beta}$
- Validation data, replication data, instrumental data

The Social Science Approach to Latent (Measurement Error) Models

- **The measurement model (one factor)**

$$
\begin{aligned}
X_1 &= \lambda_1 \xi + \delta_1 \\
X_2 &= \lambda_2 \xi + \delta_2 \\
&\;\;\vdots \qquad \vdots \qquad \vdots \\
X_p &= \lambda_p \xi + \delta_p
\end{aligned}
$$

where the λ's are loadings

- Identification: set $\lambda_1=1$ or set $\sigma^2(\xi)=1$

- Note: Even if $\lambda_1=1$, $X_2, \ldots, X_p$ are not conditionally unbiased for ξ, i.e., NOT replicates.

The Social Science Approach to Latent (Measurement Error) Models

- **The structural model**
  - one (or more) linear equations

$$\eta = \beta_0 + \beta_1\xi_1 + \beta_2\xi_2 + \beta_3 Z + \zeta$$

  - link with measurement models for η and ξ, using manifest variables ***Y, X***$_1$ and ***X***$_2$.
  - Can do simultaneous estimation of structural and measurement model parameters using SEM methods (ML, GLS, other).

One Slide Summary of SEM Methods

- Basic version (Joreskog, 1972) used ML, later GLS
- Consistent parameter and s.e. estimates if ***Y*** and ***X***'s normal
- For non-normal ***Y*** and ***X***'s, parameters consistent, s.e.'s not
- ADF, PML and DWLS versions yield consistent s.e.'s
  - the latter two with some loss of efficiency
- For discrete data, main issue is skewness and kurtosis, not the discreteness itself, provided # categories is 5 or more
- Discrete SEM (Muthen, 1984) models discreteness directly
- *Some* complex survey facilities available

Latent Variable Modeling in Practice

- Some data analysts use simultaneous SEM methods
- Many do not
  - Lack of familiarity
  - Concerns re normality “requirements”
  - Convergence problems when *N* is small, or model is large
- Alternative methods
  - OLS regression with latent variable ‘scores’ (Two-step)
  - Two-stage / instrumental variables regression (Bollen, 1996)
  - Partial Least Squares (PLS; Wold, 1982, 1985)
- Will focus on two-step and 2SLS/IV methods

Two-Step Methods

- Consider one or more LV's in a single linear equation

$$\eta = \beta_1 \xi_1 + \beta_2 \xi_2 + \beta_3 Z + \zeta \qquad (1)$$

with ζ independent of ξ_1 and ξ_2, Z exactly measured, and $E(\eta) = E(\xi_i) = E(Z) = 0$.

- Assume that ξ_i and η satisfy measurement models

$$\boldsymbol{Y} = \lambda_\eta \eta + \varepsilon \qquad \text{and} \qquad \boldsymbol{X}_i = \lambda_{\xi i} \xi_i + \delta_i, \quad i = 1, r$$

- Estimate (predict) ξ_i and η using measurement models
  - replace ξ_i and η in (1) by their predicted values
  - estimate β's by OLS

Predictors

- Factor scores $\hat{\eta} = \boldsymbol{\omega}_{\eta}^{/}\boldsymbol{Y}, \quad \hat{\xi}_i = \boldsymbol{\omega}_{i\xi}^{/}\boldsymbol{X}_i$

  - Regression $\boldsymbol{\omega}_{\eta}^{/} = \boldsymbol{\lambda}_{\eta}^{/}[\boldsymbol{\lambda}_{\eta}\boldsymbol{\lambda}_{\eta}^{/} + \boldsymbol{\theta}_{\varepsilon}]^{-1}$

  - Bartlett $\boldsymbol{\omega}_{\eta}^{/} = [\boldsymbol{\lambda}_{\eta}^{/}\boldsymbol{\theta}_{\varepsilon}^{-1}\boldsymbol{\lambda}_{\eta}]^{-1}\boldsymbol{\lambda}_{\eta}^{/}\boldsymbol{\theta}_{\varepsilon}^{-1}$

- CTT scores $\hat{\eta} = \sum Y_j / n_Y \quad , \quad \boldsymbol{\omega}_{\eta}^{/} = \boldsymbol{1}' / n_Y$
  - use when Cronbach's alpha ≥ 0.8

- IRT scores
  – nonlinear functions of ***Y***, ***X****i* – next slide

- IRT Scores

- Item characteristic function (binary manifest variables)

$$P(Y_j=1|\eta) = P_j(\eta) = \Phi[a_j(\eta - b_j)]$$

- Joint conditional probability of $\boldsymbol{Y}$

$$P(\boldsymbol{Y} \,|\eta\, ;\boldsymbol{a},\boldsymbol{b}) = \Pi_j\, [P_j(\eta\,)]^{Y_j}[1 - P_j(\eta\,)]^{1-Y_j}$$

- Assumed (or empirical) distribution of η *is* $g(\eta)$
  - hence $P(\eta \mid \boldsymbol{Y})$
  - hence $E(\eta \mid \boldsymbol{Y}) = \hat{\eta}$

- Above referred to as EAP; also have MLE and WLE

Consistency of the Two-Step Method

$$\eta = \beta' \xi + \zeta,$$

ξ and ζ independent, with zero means.

$$\hat{\eta} = \beta' \hat{\xi} + u$$

where

$$u = \zeta + \beta' (\xi - \hat{\xi}) - (\eta - \hat{\eta})$$

OLS gives a consistent estimate $\hat{\beta}$ if

$$E[\zeta + \beta' (\xi - \hat{\xi}) - (\eta - \hat{\eta})]\hat{\xi}' = 0$$

Some Sufficient Consistency Conditions

Given

- ξ and ζ independent,
- ***Y*** satisfies a trait model with univariate trait η,
- $\boldsymbol{X}_1 \dots \boldsymbol{X}_r$ satisfies a trait model with multivariate traits $\xi_1, \dots \xi_r$, and

(1) $\hat{\xi} = E(\xi \mid \boldsymbol{X})$, a calibration estimator

(2) $E(\hat{\eta}|\eta)=\eta$, i.e., $\hat{\eta}$ is conditionally unbiased.,

then $\hat{\beta}$ is consistent for β.

Example 1 $\eta = \beta' \xi + \zeta$

- Continuous manifest variables, ***Y*** and ***X***
- $\hat{\eta}$ obtained via Bartlett factor scoring
- $\hat{\xi}$ obtained via blockwise regression factor scoring

  - Bartlett scores are conditionally unbiased for η

  - Regression scores are Bayes predictors for multi-normal ***X*** and ξ

- See Skrondall and Laake (2001, Psychometrika)

Example 2 Simple Regression, Exact on Latent

$$Y = \beta\xi + \zeta$$

- Continuous Y (i.e., exactly measured η)
- Single ξ
- Discrete $\boldsymbol{X}$
- $\hat{\xi}$ obtained via IRT / EAP scoring (EB)

Percent Bias, Exact on Latent (EAP)

$$\beta = \sqrt{\rho_S^2} = .707; \quad N = 300; \quad \lambda_j^X = \sqrt{\rho_M^2} = .707$$

# cats.	# items	***B(β)***	***B(R²)***	***B(se)***
2	5	-3	-37	-2
	10	-2	-23	-3
3	5	0	-23	-4
	10	0	-14	-6
5	5	0	-20	-1
	10	0	-11	-5

Example 3 Exact on Multiple Latent

$$Y = \beta' \xi + \zeta, \qquad \xi = (\xi_1, \ldots, \xi_r)'$$

- Scores for ξ via multivariate (blockwise) EAP scoring
  - $\tilde{\xi} = E(\xi \mid \boldsymbol{X}_1, \ldots, \boldsymbol{X}_r)$
  - not the usual case

- Typically, univariate (factorwise) scoring is used, i.e.,
  - $\hat{\xi}_k = E(\xi_k \mid \boldsymbol{X}_k)$, $k = 1, \ldots, r$
  - yields estimate $\hat{\beta}$
- $\hat{\beta}$ consistent if ξ_k's are independent (see also S and L)

Percent Bias, Exact on Latent (univ. EAP)

$(\rho_R^2 = 0.5;\ \ \gamma_1 = \gamma_2;\ \ N = 300;\ n = 5;\ \ \lambda_j^{X_1} = \lambda_j^{X_2} = .707)$

# cats.	$\rho(\xi_1, \xi_2)$	$B(\beta)$	$B(R^2)$	B(se)
2	0.5	9	-27	0
2	0	-2	-35	-2
2	-0.5	-27	-52	-1
5	0.5	7 (0)	-14 (-13)	-1(-4)
5	0	0	-19	0
5	-0.5	-16	-32	0

Multivariate EAP

- **Note 1**

$$\eta = \beta' \xi + \gamma Z + \zeta$$

If Z is correlated with ξ, then neither univariate nor multivariate EAP scoring of ξ will yield consistent parameter estimates.

Need $\hat{\xi} = E(\xi \mid X, Z)$, **or a simpler alternative**.

- **Note 2**

There is no cond'lly unbiased IRT scoring procedure

- Bias (EAP) = O(1/n)
- Bias (WLE) = o(1/n)

WLE / mult EAP $\equiv$ Bartlett / Regression?

Analysis of Bias in Two-Step Regression (Croon, 2002; C and R, 2002)

- Continuous case / discrete case not tractable
- Simple regression

$$\eta = \beta\xi + \zeta$$

$$\boldsymbol{Y} = \lambda_\eta \eta + \boldsymbol{\varepsilon} \quad \text{and} \quad \boldsymbol{X} = \lambda_\xi \xi + \boldsymbol{\delta}$$

$$\hat{\eta} = \boldsymbol{\omega}_\eta^{/} \boldsymbol{Y} \quad \text{and} \quad \hat{\xi} = \boldsymbol{\omega}_\xi^{/} \boldsymbol{X}$$

- OLS $\quad \beta^* = E(\hat{\xi}\hat{\eta}) / E(\hat{\xi}^2) = \beta\rho_\xi^2 \left(\dfrac{\pi^\eta}{\pi^\xi} \right)$

where $\pi^\eta = \boldsymbol{\omega}_\eta^{/} \boldsymbol{\lambda}_\eta$ and $\pi^\xi = \boldsymbol{\omega}_\xi^{/} \boldsymbol{\lambda}_\xi$

Analysis of Two-Step Bias, cont'd

$$\beta^* = \beta\rho_\xi^2\left(\frac{\pi^\eta}{\pi^\xi}\right)$$

- $\hat{\eta}$ and $\hat{\xi}$ conditionally unbiased when $\pi^\eta = \pi^\xi = 1$
  - yields the classical attenuation result

- Otherwise, error variance not only source of bias
  - $\pi^\eta \neq 1$ implies bias if η regressed on exact $\xi = X$

- Can recover S and L result
  - Bartlett scoring $\pi^\eta = 1$
  - Regression scoring $\pi^\xi = \rho_\xi^2$

Analysis of Two Step Bias, cont'd

- Also, for $\sigma_\eta^2 = \sigma_\xi^2 = 1$

$$N^{1/2} V^{1/2}(\hat{\beta}^*) \to \left(\frac{\pi^\eta}{\pi^\xi}\right)\left(\frac{\rho_\xi^2}{\rho_\eta^2}\right)^{1/2} \left(1 - \beta^2 \rho_\xi^2 \rho_\eta^2\right)^{1/2}$$

- Thus, for a test of H_0: $\beta = 0$, π^η / π^ξ cancels
  - test *power* depends only on ρ_η^2 and ρ_ξ^2 .
  - i.e., on error variance not prediction bias

- OLS-based estimate of $V^{1/2}(\hat{\beta}^*)$ also consistent

- **No equivalent results for discrete / IRT two-step**
  - above is a rough guide

Percent Bias for Latent (EAP) on Exact

($\gamma = \sqrt{\rho_S^2} = .707$; N = 300; $\lambda_i^Y = \sqrt{\rho_M^2} = .707$)

# cats.	# itms	$B(\beta)$		$B(R^2)$		$B(se)$	
		IRT	CTT	IRT	CTT	IRT	CTT
2	5	-31	-43	-31	-32	15	18
	20	-11	-43	-11	-13	7	18
5	5	-19	-32	-19	-20	0	4
	20	-5	-32	-6	-6	-3	2

Percent Bias for Exact on Latent (EAP)

($\gamma = \sqrt{\rho_S^2} = .707$; N = 300; $\lambda_i^Y = \sqrt{\rho_M^2} = .707$)

# cats.	# itms	$B(\beta)$		$B(R^2)$		$B(se)$	
		IRT	CTT	IRT	CTT	IRT	CTT
2	5	-3	17	-37	-37	-2	-1
	20	-1	53	-13	-15	-2	-2
5	5	0	20	-20	-20	-1	-1
	20	0	40	-6	-6	-4	-4

Notes for Latent on Latent Two Step

- For latent on latent two-step regression, with similar measurement models for response and explanatory latent variables, CTT and IRT/EAP biases are similar

- Effects of category numbers and scale lengths are similar for simple and multiple regression, though degree of correlation between latent explanatory variables affects magnitudes of biases

- To minimize bias, use long scales and a minimum of 5 categories per item

$\hat{\beta}$

Instrumental Variables / Two Stage Least Squares

$$Y = \beta' \xi + \zeta, \qquad Y \in R, \ \xi \in R^r$$

$$\boldsymbol{X} = \xi + \delta, \qquad \xi, \delta \text{ and } \zeta \text{ independent and normally distributed}$$

- Assume $Z \in R^q$ such that $E(Z\xi') \neq 0$,

  but $\boldsymbol{Z}$ is uncorrelated with δ *and* ζ.
- Let $\hat{X} = AZ$, and write

$$Y = \beta' \hat{X} + u, \qquad u = \beta'(\boldsymbol{X} - \hat{X}) + (\zeta - \beta'\delta)$$

- Then, u is uncorrelated with $\hat{X} = AZ$, and $\boldsymbol{Z}$, if

  $E[(\mathbf{X} - \boldsymbol{AZ})\boldsymbol{Z}'] = \mathbf{0} \Rightarrow \boldsymbol{A} = E(\boldsymbol{XZ}')E^{-1}(\boldsymbol{ZZ}')$ 1st stage
- Regress Y on $\boldsymbol{Z}$ to get estimate of $\beta'\boldsymbol{A} = \boldsymbol{G}$
- Hence, from $\beta' = \boldsymbol{GA}'(\boldsymbol{AA}')^{-1}$, get estimate $\hat{\beta}$ 2nd stage

Bollen's (1996) 2SLS Approach (see also Joreskog and Sorbom, 1993)

$$\eta = \beta_0 + \beta_1 \xi_1 + \beta_2 \xi_2 + \zeta$$

Measurement Models

$$
\begin{aligned}
Y_1 &= \eta + \varepsilon_1 \\
Y_2 &= \lambda_{\eta 2} \eta + \varepsilon_2 \\
&\vdots \\
Y_p &= \lambda_{\eta p} \eta + \varepsilon_p
\end{aligned}
$$

$$
\begin{aligned}
X_{i1} &= \xi_i + \delta_{i1} \\
X_{i2} &= \lambda_{\xi i2} \xi_i + \delta_{i2} \\
&\vdots \\
X_{ip} &= \lambda_{\xi\, ip} \xi_i + \delta_{ip}
\end{aligned}
$$

i = 1, 2

Hence

$$(Y_1 - \varepsilon_1) = \beta_0 + \beta_1 (X_{11} - \delta_{11}) + \beta_2 (X_{21} - \delta_{21}) + \zeta$$

i.e., $$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + u ,$$

where $$u = \zeta - \beta_1 \delta_{11} - \beta_2 \delta_{21} + \varepsilon_1$$

2SLS Approach (cont.)

Instruments : $X_{i2}, \ldots, X_{ip}$, for $i = 1, 2$

First Stage:

Regress X_{11} on $q = 2(p-1)$ instruments to get $\hat{X}_{11}$

Regress X_{21} on q instruments to get $\hat{X}_{21}$

Second Stage:

OLS of Y_1 on $\hat{X}_{11}$ and $\hat{X}_{21}$ to get consistent estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$

Issues

- **Bias**
  - increases as # instruments increases
  - for univariate ξ, under the Bollen setup

$$\hat{\beta} - \beta \propto [\, r(\, p-2\,) - 1] \,/\, N$$

  - similar to Nagar (1959) for individual equations in econometric systems
- **Finite Sample Moments**
  - *may* only exist up to degree $r(p-2)$
  - ref Mariano (1972), Phillips (1983) for discussion of limited info estimates in econometric systems
- **Tradeoff?**

Bias vs Scale Length, For Two Explanatory LVs

True R^2 =0.5; $\rho(\xi_1,\xi_2)$=0.5; λ's =0.775; 5 cats; N=150

# Items / scale (*p*)	**# Free IV's** $q\text{-}r = r(p\text{-}2)$	**B(b)**	**B(R^2)**
2	0	-0.8	1.5
3	2	0.0	0.0
5	6	-1.5	-3.0
10	16	-4.9	-9.6
20	36	-11.0	-20.7

Issues, continued

- $Y_2, \ldots, Y_p$ not used in the single equation application of Bollen's method

  - select the one with the highest loading?

  - the highest R^2 on the predictor *X*'s ?

  - Use Bartlett (conditionally unbiased) score?

  - other?

2SLS Simulation, Two Explanatory LVs, Discrete Indicators

- Sample Size = 300, β's=.707, λ's = .775

# items	Method	%B(β)	%B(R^2)
5Y, 5X1, 5X2 (C's α = 0.88)	2SLS(A)	0.2	0.5
	2SLS(B)	-0.2	-0.4
	2-Step (CTT)	-9.7	-18.4
	2-Step (IRT)	-9.8	-19.7
	Discrete-SEM	1.2	2.6
10Y. 10X1, 10X2 (C's α = 0.94)	2SLS(A)	0.1	0.2
	2SLS(B)	-0.5	-0.6
	2-Step (CTT)	-4.9	-9.5
	2-Step (IRT)	-5.1	-9.9
	Discrete-SEM	0.4	1.0

- **2SLS(A): Item Y_1 used for $\hat{\eta}$. 2SLS(B): Bartlett score for $\hat{\eta}$.**
- **2 Instruments: means of $X_{1,2}$ … $x_{1,10}$, and $X_{2,2}$ …$X_{2,10}$, resp.**

2SLS for Probit with Latent Predictors

(Bollen, Thomas, Wang & Hipp, 2005; SAMSI / NPCDS)

$$P\,(\,Y = 1 \mid \boldsymbol{\xi}\,,\, \boldsymbol{\beta}\,) = f\,(\boldsymbol{\xi}\,,\, \boldsymbol{\beta}\,)$$

Consider probit model:

$$y^* = \alpha + \sum_{i=1}^{r} \beta_i \xi_i + \zeta$$

with $\xi \sim N(\,\mathbf{0}\,,\, \boldsymbol{\Sigma}_{\xi}\,)$, independent of $\zeta \sim N(0,\, \mathbf{1}\,)$

and $\mathrm{Y} = 1$ when $\mathrm{y}^* > 0$ and $\mathrm{Y} = 0$ when $\mathrm{y} \leq 0$

Measurement models as before for $\xi_1, \ldots \xi_r$

Hence unbiased predictors $X_{1,1}, \ldots, X_{r,1}$ *for* $\xi_1, \ldots \xi_r$

Assume instruments $Z_1, \ldots, Z_q$ $(q \geq r)$

2SLS Probit with Latent Predictors, cont'd

Step 1. Regress X_{i1}'s in turn on Z's to get $\hat{X}_{11}, \hat{X}_{21}, \ldots, \hat{X}_{k1}$

(Variance of disturbance now no longer ≠ 1)

Step 2. Probit regression of Y on $\hat{X}_{11}, \hat{X}_{21}, \ldots, \hat{X}_{k1}$ to get $\hat{\beta}_1, \ldots, \hat{\beta}_k$

Step 3. Estimate variance of disturbance term – need polychoric

correlation of y* and **X** – obtain bias correction factor

Step 4. Correct $\hat{\beta}_1, \ldots, \hat{\beta}_k$ to get $\hat{\beta}_1^*, \ldots, \hat{\beta}_k^*$

Step 5. Get linearization estimator using plug-in estimate of C(y*, **X**)

Note: $\hat{\beta}^*$ are consistent, unlike the calibration method of CRS.

2SLS Probit with Latent Predictors, cont'd

Simulation

3 latent predictors, R^2 = 0.5; 3 indicators per latent predictor;

CD = 0.5 for each indicator; 2 IV's per latent predictor;

Sample Size (N)	$\overline{\mathrm{B}}$	$\sigma(\overline{\mathrm{B}})$	Bias Detected? $(\lvert\overline{\mathrm{B}}\rvert > 2\sigma(\overline{\mathrm{B}}))$
100	.0311	.0091	Yes
200	.0187	.0046	Yes
500	.0092	.0026	Yes
1000	.0018	.0017	No

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

THANKS FOR YOUR ATTENTION