

Catalogue no. 11-522-XIE

**Statistics Canada International
Symposium Series - Proceedings**

**Symposium 2005 :
Methodological Challenges for
Future Information needs**



2005



**Statistics
Canada**

**Statistique
Canada**

Canada

TWO-STEP REGRESSION WITH LATENT VARIABLES, REVISITED

D. R. Anderson¹ and R. L. Marshall²

ABSTRACT

This paper reviews the use of two-step regression with latent variables. We discuss the advantages and disadvantages of this approach, and compare it to other methods. We also discuss the use of two-step regression with latent variables in the context of structural equation modeling (SEM).

KEY WORDS: Latent variables, IRT, two-step regression, OLS, structural equation modeling, SEM

1. INTRODUCTION

1.1 Latent Variable Regression

This paper reviews the use of two-step regression with latent variables. We discuss the advantages and disadvantages of this approach, and compare it to other methods. We also discuss the use of two-step regression with latent variables in the context of structural equation modeling (SEM).

¹ Scott Beatty, University of Alberta, Canada (sbeatty@ualberta.ca)
² Scott Anderson, York University, Toronto, Canada (scott.anderson@yorku.ca)

3. TWO-STEP METHODS

Let η and ξ be random variables with joint distribution $f(\eta, \xi)$. Let $\bar{\xi}$ be the sample mean of ξ . The OLS estimator of η is given by

$$\hat{\eta} = \gamma' \bar{\xi} + u, \quad (4)$$

$$u = \zeta + \gamma'(\xi - \bar{\xi}) - (\eta - \eta). \quad (5)$$

OLS estimator of η is given by (4) and $\bar{\xi}$ is given by

$$E\bar{\xi}[\zeta + (\xi - \bar{\xi})'\bar{\gamma} - (\eta - \eta)] = 0, \quad (6)$$

because $E(u) = 0$. This can be written as

$$E(\bar{\xi}\eta) = E(\bar{\xi}\bar{\xi}')\bar{\gamma}. \quad (7)$$

The OLS estimator of η is given by (7) and $\bar{\xi}$ is given by

3.1 Sufficient Conditions for OLS Consistency

Consistency of OLS estimator of η requires the following conditions:

- (i) The random variables X, ξ are independent.
- (ii) The random variables X, ξ are independent of Y .
- (iii) $E(\eta) = \eta$.
- (iv) $E(\xi) = \xi$.
- (v) $\bar{\xi} = (\xi_1, \dots, \xi_n)'$ is a consistent estimator of ξ .
- (vi) $E_Y(\eta) = \eta$, $E(\xi) = \xi$.

Proof: Let η and ξ be random variables with joint distribution $f(\eta, \xi)$. Let $\bar{\xi}$ be the sample mean of ξ . The OLS estimator of η is given by

3.2 OLS Regression of a Measured Y on a Single Latent Predictor

Let Y be a measured variable and η be a latent variable. Let ξ be a latent variable. The OLS estimator of η is given by

3.3 OLS Regression of a Measured Y on Multiple Latent Predictors

We let $r > 1$, and let ξ_k be a vector of latent predictors. Let Y be a measured variable. The OLS regression of Y on ξ_k is given by
$$Y = \gamma + \sum_{k=1}^r \beta_k \xi_k + \eta \quad (8)$$
 where η is the error term. The OLS estimates of γ and β_k are given by
$$\hat{\gamma} = E(Y - \sum_{k=1}^r \hat{\beta}_k \hat{\xi}_k) \quad (9)$$
 where $\hat{\xi}_k = E(\xi_k | X)$, $k=1, \dots, r$.

Here, $\xi_k \neq \hat{\xi}_k$, and the OLS estimates of γ and β_k are biased. The bias is given by
$$E(\hat{\gamma}) - \gamma = -\sum_{k=1}^r \beta_k E(\xi_k - \hat{\xi}_k) \quad (10)$$

3.4 OLS Regression with Factor Scores

The bias in the OLS regression of Y on ξ_k is given by
$$E(\hat{\gamma}) - \gamma = -\sum_{k=1}^r \beta_k E(\xi_k - \hat{\xi}_k) \quad (11)$$
 where $\hat{\xi}_k = E(\xi_k | X)$. The bias is zero if $E(\xi_k - \hat{\xi}_k) = 0$, which is the case if ξ_k is a linear function of X .

3.5 Bias in the General Case of IRT-Based Latent Regression

The bias in the IRT-based latent regression is given by
$$E(\hat{\gamma}) - \gamma = -\sum_{k=1}^r \beta_k E(\xi_k - \hat{\xi}_k) \quad (12)$$
 where $\hat{\xi}_k = E(\xi_k | X)$.

4. COMPARISON METHODS

4.1 Two-Step Regression with CTT Scores

Cornwell (2002) proposed a two-step regression method. The first step is to regress Y on X to get
$$Y = \bar{\lambda}_\eta \eta + \bar{\epsilon} \quad (13)$$
 where $\bar{\lambda}_\eta = E(\eta | X)$ and $\bar{\epsilon} = Y - \bar{\lambda}_\eta \eta$. The second step is to regress $\bar{\epsilon}$ on X to get
$$\bar{\epsilon} = \bar{\lambda}_\xi \xi + \bar{\delta}_i \quad (14)$$
 where $\bar{\lambda}_\xi = E(\xi | X)$ and $\bar{\delta}_i = \bar{\epsilon} - \bar{\lambda}_\xi \xi$.

The bias in the two-step regression is given by
$$E(\hat{\gamma}) - \gamma = -\sum_{k=1}^r \beta_k E(\xi_k - \bar{\lambda}_\xi \xi) \quad (15)$$
 where $\bar{\lambda}_\xi = E(\xi | X)$. The bias is zero if $E(\xi_k - \bar{\lambda}_\xi \xi) = 0$, which is the case if ξ_k is a linear function of X .

4.2 Two-Stage Least Squares for Latent Regression

Bollen (1996) proposed a two-stage least squares method. The first step is to regress Y on X to get
$$Y = \lambda_\eta \eta + \epsilon \quad (16)$$
 where $\lambda_\eta = E(\eta | X)$ and $\epsilon = Y - \lambda_\eta \eta$. The second step is to regress ϵ on X to get
$$\epsilon = \lambda_\xi \xi + \delta_i \quad (17)$$
 where $\lambda_\xi = E(\xi | X)$ and $\delta_i = \epsilon - \lambda_\xi \xi$.

(1). A - e e e , e a c e e d a e a c e a e d e e d c . T e - a e e d d e e - e e d a e a a a b e a e e e c e d a a e c e a e d e e d c b c e a e d e e d a , a d c a e e d b a e a e a a a e e e b a a e e d a . T e e a e e e c e d e e a e e e e c e a e e e c e e . T e e d e a a e d b e e e e d c e e e e a e , a d e e e e e a a b e e e d e d c e e c d a e . T e c e a a e e e a e .

4.3 Discrete SEM Estimation

SEM e a c e c b e a a e e e d e c a E a (1) e a e e d e e e e c e d E a (9). D c e e SEM e a b a e d a e e e SEM d e c e e e c Y a d X_i, i=1,...,r, a e e e e e a d e d a a e , e d c e e c e a b a e d b c a e e a e e a c c d e d e e e a e d e d a a . D c e e SEM e a e e e d a SEM c e a c a e , a d e a e a e e e b a e d Mplus. D c e e SEM e a e a e c e .

5. SIMULATION

5.1 Simulation Design

T e d e c e a b a e d e c e a d e c b e d a b e d c e e SEM. C e e d e a a e e b L (2004). T e e e a e a e e e a e d : (a) e e a a a e a a e a a b e e a a c e e c a d e (1), c a E(η) = E(ξ) = 0 a d σ_η² = σ_ξ² = 1; (b) e e a a a e a e e e , y_j a d x_j, c e b d e (1) a d e e a e e d e (9); (c) a a y_j a d x_j d c e e e c e e d a a e e d e d e e a e b e c a d - e c c a e e ; a d (d) e CTT c a e , e - a d a d a e e d c e e e c e a e e a e a d a a c e e , e e a d e e e d a a c e . T e a e d e c b e d b e a e e e e c a e e a IRT-EAP c , a d a d e a b e c a a a e e b a e b e e e - e , - a e a d d c e e SEM a a c e . T e e b a d c a a e d a e d c a e a e e d b S d a a d L a a e (2001), a e e e c e a e e a e b a e e a e e c a e e a a e e (), e c e c e d e e a e c a d e , ρ_R², a d e OLS e a e e a a e e a d a d e (). F a , e e a e d e d a

$$B_{\gamma} = 100 \left[\frac{E(\gamma)}{\gamma} - 1 \right], B_{R^2} = 100 \left[\frac{E(R^2)}{\rho_R^2} - 1 \right], \text{ a d } B_{SE} = 100 \left[\frac{E(SE)}{V^{1/2}(\gamma)} - 1 \right], \quad (10)$$

e e E[·] a d V[·] d e e e c a e e c e d a e a d a a c e b a e d 500 a e c a .

5.2 Exact Response Variable and a Single Latent Explanatory Variable, with EAP Scoring

T a b e 1 d a a e e - e e e a e a c e e a a b e a e a e e a a a a b e c e d b e a IRT-EAP c e . T e a e d c e c e d e e a b e e a e e e e a (ρ_R²) a d e a c d d a e e a e e e a (ρ_M²) e e e e a 1/2. T e b e c a e e e e (m+1) a d e b e e e c a e (n) d c a e d e c e e a b e . T e a e e , N , a e a 300, a d a e e e e e a e d a c c d a e c c a e d b . F e d c e a b e c a b e e e a e e e c a e d c a e b e , a a e e b a b e e (e e a e e c e a e) 3 a d 5 c a e e e e e , a d c e e e b a e . T e c a e R² e a e e c e c e d e e a e e e e e a e b a e d . T c e e e , c e EAP c d e e a e e e a a c e e a e a a b e . I

ca a be ee e abe a eba R^2 dec ea e a e be e cea e (a a e be ca e e cea e), a a c e e c edc a e ea e e e a a ce EAP c e e e a $n \rightarrow \infty$. Fa , e a c Table 1 ca be ee a ba e e a da d OLS e ae a a ee a da de ae a .

Table 1

Percent Relative Biases for Exact on a Single Latent: IRT-EAP Scoring

$$(N = 300; \gamma = \sqrt{\rho_R^2} = .707; \lambda_j^X = \sqrt{\rho_M^2} = .707)$$

$m + 1$	n	B_γ	B_{R^2}	B_{SE}
2	5	-3	-37	-2
	10	-2	-23	-3
	20	-1	-13	-2
3	5	0	-23	-4
	10	0	-14	-6
	20	0	-8	-7
5	5	0	-20	-1
	10	0	-11	-5
	20	0	-6	-4

Note. $m+1 =$ be e ca e e ; $n =$ ca e e ; $N =$ a e e .

5.3 Exact Response Variable and Two Latent Explanatory Variables: Univariate and Multivariate IRT-EAP Scoring

Table 2 d a e a e a - e e e a e ac e e a abe a e edc a abe a dc a e a a ee ba e ba ed a ae IRT-EAP c e a d a ae IRT-EAP c e . Pa a ee e ae e a ea e Table 1, e ce a a e c e d 5- e ca e , a d e ce e add ac ea be ee e ae edc ξ_1 a d ξ_2 . Te da d c Table 2 ba e a ae IRT-EAP c . A ed c ed ea e , a a ee e ae be ba ed ca e e ec ea be ee ξ_1 a d ξ_2 e , a d edc cea b e . Ne a ba e e e e a a ee a e a e a ed Table 2. C 5 a d 6 Table 2 ee e a e a ae IRT-EAP c e . Tee ee ba ed *Mplus* (M a d M , 2001), a d e dd e c e e a eb a ca e . I ca be ee a a a ee ba e ae e e be 5-ca e e , e ec e ec ea , a ed c ed b e . A be e, R^2 ba ed e ec e e e EAP c e ed.

5.4 EAP Two-Step Estimation Versus The Identified Alternatives

Table 3 de a c a a a ee a d R^2 ba e d : (1) - e e e ea EAP c e ; (2) - e e e ea CTT c e ; (3) - a e ea ae (2SLS);

Table 2

Percent Relative Biases for Exact on Two Latents: Univariate and Multivariate IRT-EAP Scoring

$$(N = 300; \rho_R^2 = 0.5; \gamma_1 = \gamma_2; n = 5; \lambda_{\xi_1} = \lambda_{\xi_2} = 0.707)$$

$m + 1$	$\rho(\xi_1, \xi_2)$	U a a e EAP		M a a e EAP	
		$B(\gamma)$	$B(R^2)$	$B(\gamma)$	$B(\gamma)$

2	0.5	9	-27	--	--
2	0.0	-2	-35	--	--
2	-0.5	-27	-52	--	--
5	0.5	7	-14	0	-13
5	0.0	0	-19	0	-19
5	-0.5	-16	-32	-2	-32

and (4) discrete SEM estimation. The average bias of the parameter estimates is reported in Table 2, where the bias is measured as the difference between the true parameter value and the estimated value. The IRT-EAP estimator shows the lowest bias across all parameters, followed by the EAP estimator. The bias of the 2SLS estimator is generally larger than that of the other two estimators.

In addition, Table 3 reports the bias of the variance-covariance matrix estimates. The bias of the variance-covariance matrix estimates is measured as the difference between the true variance-covariance matrix and the estimated variance-covariance matrix. The bias of the variance-covariance matrix estimates is generally smaller than that of the parameter estimates. The bias of the 2SLS estimator is generally larger than that of the other two estimators.

Table 3

Bias Comparisons for Two-Step, Two-Stage and Discrete SEM Estimation
 ($N = 300$; $\rho_R^2 = 0.5$; $\gamma_1 = \gamma_2$; $\rho(\xi_1, \xi_2) = 0.5$; $n = 5$; $\lambda_\eta = \lambda_{\xi_1} = \lambda_{\xi_2} = 0.775$)

Number of Indicators	Method	$B(\gamma)$	$B(R^2)$
5Y, 5X ₁ , 5X ₂	2-Step (EAP)	-9.8	-19.7
	2-Step (CTT)	-9.7	-18.4
	2SLS	0.2	0.5
	Discrete-SEM	1.2	2.6
10Y, 10X ₁ , 10X ₂	2-Step (EAP)	-5.1	-9.9
	2-Step (CTT)	-4.9	-9.5
	2SLS	0.1	0.2
	Discrete-SEM	0.4	1.0

6. CONCLUSIONS AND RECOMMENDATIONS

The results of this study show that the IRT-EAP estimator is the most accurate and unbiased estimator of the parameters of a two-stage, two-step structural equation model. The bias of the IRT-EAP estimator is generally smaller than that of the other two estimators. The bias of the variance-covariance matrix estimates is generally smaller than that of the parameter estimates. The bias of the 2SLS estimator is generally larger than that of the other two estimators.

The results of this study also show that the IRT-EAP estimator is the most accurate and unbiased estimator of the variance-covariance matrix. The bias of the variance-covariance matrix estimates is generally smaller than that of the other two estimators. The bias of the 2SLS estimator is generally larger than that of the other two estimators.

In conclusion, the IRT-EAP estimator is the most accurate and unbiased estimator of the parameters and variance-covariance matrix of a two-stage, two-step structural equation model. The bias of the IRT-EAP estimator is generally smaller than that of the other two estimators. The bias of the variance-covariance matrix estimates is generally smaller than that of the parameter estimates. The bias of the 2SLS estimator is generally larger than that of the other two estimators.

