行政院國家科學委員會專題研究計畫成果報告

# 廣義 LISREL 模式（III）
# Generalized LISREL Models (III)

計畫類別：☑個別型計畫　☐整合型計畫

計畫編號：**NSC 88－2118－M－002－008**

執行期間：**87 年 8 月 1 日至 88 年 7 月 31 日**

個別型計畫：計畫主持人：**胡賦強 助理教授**
共同主持人：

整合型計畫：總計畫主持人：
子計畫主持人：

註：整合型計畫總報告與子計畫成果報告請分開編印各成一冊，彙整一起繳送國科會。

處理方式：☐可立即對外提供參考
（請打✓）☑一年後可對外提供參考
☐兩年後可對外提供參考
（必要時，本會得展延發表時限）

執行單位：**國立台灣大學 公共衛生學院 流行病學研究所**
**生物醫學統計組**

中華民國　89　年　3　月　20　日

Report

# Generalized Path Analysis for Recursive Models (II): An Instrumental Variable Method

**Fu-Chang Hu, Tien-Lung Tsai, and Wen-Yi Shau**[1]

Division of Biostatistics
Graduate Institute of Epidemiology
College of Public Health
National Taiwan University
Taipei, Taiwan 100
R.O.C.

March 20, 2000

[1]Correspondence: Fu-Chang Hu, Division of Biostatistics, Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, 1 Jen-Ai Road, Section 1, Room 1541, Taipei, Taiwan 100, R.O.C., Phone: +886 2 / 2397-0800 × 8349 or +886 2 / 2394-2050, Fax: +886 2 / 2351-1955, Email: fchu@ha.mc.ntu.edu.tw. The first author wishes to thank Professor Karl Jöreskog for introducing him to the use of the LISREL package (Version 8.0) in an invited workshop held in Taipei, Taiwan, R.O.C., August 29-31, 1995. This research was financially supported by the National Science Council of the Republic of China (NSC 86-2115-M-002-016, NSC 87-2118-M-002-003, and NSC 88-2118-M-002-008).

# Abstract

The statistical models for a system of linear regression-like structural equations with observed variables, which include *simultaneous equations model* (SiEM) and *path analysis* (PA) model, have been used extensively for exploring or examining the plausible causal relationship among several continuous response variables by economists and social scientists. These two classes of statistical models are essentially the same except that their estimation methods differ. A system without any reciprocal effects between the response variables of the structural equations is called a *recursive* model, which is particularly useful in analyzing longitudinal data due to the temporal order of the response variables. When all the error terms of the structural equations in a recursive SiEM or PA model are mutually independent, it is called the *fully* recursive model, for which consistent estimates of the structural coefficients can be obtained equation-by-equation separately. A recursive model with correlated error terms between some of the structural equations is called the *partially* recursive model, for which the equation-by-equation approach is usually not valid and the estimation of the structural coefficients should be based on the whole system of equations. In this study, we generalize the standard SiEM and PA method to the situations in which the outcomes of a partially recursive system are a mixture of discrete and continuous variables. Specifically, we combine the *indirect least squares* (ILS), *two-stage least squares* (2SLS), and *instrumental variable* (IV) estimation methods of SiEM with the *iterative reweighted least squares* (IRLS) algorithm of generalized linear models (GLMs) to estimate the structural coefficients of a partially recursive model with the response variables of a mixed type. Statistical properties of our estimators, especially the IV estimator, are examined. Our simulation studies show promising results.

## Keywords:

Causal analysis, Path analysis, Simultaneous equations model, Structural equation model, Recursive model, Generalized linear models, Discrete responses, Mixed responses, Instrumental variable.

# 1 Introduction

In the past 50 years or so, the *path analysis* (PA) method has been used extensively in various areas of social sciences for exploring and/or examining the plausible causal relationship among several response variables. Independently, econometricians have developed the *simultaneous equations model* (SiEM) for exploring and/or examining the plausible structural relationship among several endogenous variables. These two kinds of statistical models are essentially the same except that their estimation methods are different.

Both PA and SiEM require that all the response or endogenous variables be continuous random variables and their joint distribution be multivariate normality (or at least they are symmetrically distributed). In this project, we generalize the standard PA and SiEM to deal with binary, continuous, and/or counts response or endogenous variables as in *generalized linear models* (GLMs) and call the new models the "generalized path analysis" (GPA) models or equivalently the "generalized simultaneous equations models" (GSiEM).

# 2 Model

The *fully* recursive GPA (GSiEM) models are trivial since they can be solved equation-by-equation. To start with, we consider the following simplest *partially* recursive GPA (GSiEM)

model:

$$g_1(\mu_1) = \beta_{10} + \beta_{1x_1} X_1 + \beta_{1x_2} X_2 \quad (1)$$

$$\mu_2 = \beta_{20} + \beta_{2x_1} X_1 + \beta_{2x_3} X_3 + \gamma_{2y_1} Y_1 \quad (2)$$

where $(X_1, X_2, X_3)$ are the covariates (or exogenous variables), $\mu_1$ and $\mu_2$ are the means of the responses (or endogenous variables) $Y_1$ and $Y_2$ respectively,

$$Y_1 \sim \text{Exponential family of distributions, e.g. Binomial } (m, \mu_1),$$

$$Y_2 \sim \text{Normal } (\mu_2, \sigma_2^2),$$

and $g_1$ is the *link* function for $\mu_1$. For simplicity, we assume that the *link* function for the mean of $Y_2$ is the *identity* function, but the other link functions, e.g. *log*, or the other GLMs will be considered later.

# 3 Estimation Methods

## 3.1 Indirect Least Squares (ILS) Estimator

- **Step 1:**

  Fit the GLM for the response $Y_1$ on the covariates $X_1$ and $X_2$ using the IRLS algorithm to obtain the estimates $\widehat{\beta}_{10}$, $\widehat{\beta}_{1x_1}$, and $\widehat{\beta}_{1x_2}$ of the corresponding coefficients in the first equation.

- **Step 2:**

  In **Step 1**, the *pseudo*-response variable $Z_1$ (defined below) on the original covariates $X_1$ and $X_2$ on the convergence:

$$\begin{aligned} Z_1 &= \left(\widehat{\beta}_{10} + \widehat{\beta}_{1x_1} X_1 + \widehat{\beta}_{1x_2} X_2\right) + g_1'\left(\widehat{\mu}_1\right)\left(Y_1 - \widehat{\mu}_1\right) \\ &= g_1\left(\widehat{\mu}_1\right) + g_1'\left(\widehat{\mu}_1\right)\left(Y_1 - \widehat{\mu}_1\right) \quad \left(g_1'\left(\widehat{\mu}_1\right)\left(Y_1 - \widehat{\mu}_1\right) = \epsilon_{1,1}^*\right) \\ Y_1 &= \widehat{\mu}_1 + \frac{Z_1 - g_1\left(\widehat{\mu}_1\right)}{g_1'\left(\widehat{\mu}_1\right)} \end{aligned}$$

  where

$$\widehat{\mu}_1 = \widehat{Y}_1 = g_1^{-1}\left(\widehat{\beta}_{10} + \widehat{\beta}_{1x_1} X_1 + \widehat{\beta}_{1x_2} X_2\right).$$

- **Step 3:**

  Notice that by plugging

$$Y_1 = \widehat{Y}_1 + \frac{Z_1 - g_1\left(\widehat{\mu}_1\right)}{g_1'\left(\widehat{\mu}_1\right)}$$

  from **Step 2** into the second equation, we have

$$\begin{aligned} Y_2 &= \beta_{20} + \beta_{2x_1} X_1 + \beta_{2x_3} X_3 + \gamma_{2y_1} Y_1 + \epsilon_2 \\ &= \beta_{20} + \beta_{2x_1} X_1 + \beta_{2x_3} X_3 + \gamma_{2y_1} \widehat{Y}_1 + \left[\gamma_{2y_1}\left(\frac{Z_1 - g_1(\widehat{\mu}_1)}{g_1'(\widehat{\mu}_1)}\right) + \epsilon_2\right] \\ &= \beta_{20} + \beta_{2x_1} X_1 + \beta_{2x_3} X_3 + \gamma_{2y_1} \widehat{Y}_1 + \epsilon_{2,1}^*. \end{aligned} \tag{4.1}$$

  1. For simplicity, we can just use the OLS method to obtain

$$\widehat{\boldsymbol{\beta}}_{2,ILS} = \left(\mathbf{X}_2^{+\mathrm{T}} \mathbf{X}_2^{+}\right)^{-1} \mathbf{X}_2^{+\mathrm{T}} \mathbf{Y}_2$$

where the design matrix is $\mathbf{X}_2^+ = \left[\mathbf{1}, \mathbf{X}_1, \mathbf{X}_3, \widehat{\boldsymbol{Y}}_1\right]$. Note that WLS estimator will be considered in a later subsection.

2. Alternatively, to gain efficiency, we can also regress $[\mathbf{Z}_1, \mathbf{Y}_2]^{\mathrm{T}}$ on $\mathbf{X}_1$ and $\mathbf{X}_2^+$ jointly as in a *regression system* (RS) or *seemingly unrelated regression* (SUR) model, i.e.

$$\begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2^+ \end{bmatrix} \mathbf{B} + \begin{bmatrix} \boldsymbol{\epsilon}_{1,1}^* \\ \boldsymbol{\epsilon}_{2,1}^* \end{bmatrix}$$

using the *iterated feasible generalized least squares* (IFGLS) method to obtain the estimate $\widehat{\mathbf{B}}$ of $\mathbf{B}$, where $\mathbf{X}_1 = [\mathbf{1}, \mathbf{X}_1, \mathbf{X}_2]$, $\mathbf{X}_2^+ = \left[\mathbf{1}, \mathbf{X}_1, \mathbf{X}_3, \widehat{\boldsymbol{Y}}_1\right]$. , and $\mathbf{B}^{\mathrm{T}} = [\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}] = [\beta_{10}, \beta_{1x_1}, \beta_{1x_2}, \beta_{20}, \beta_{2x_1}, \beta_{2x_3}, \gamma_{2y_1}]$.

## 3.2 Two-Stage Least Squares (2SLS) Estimator

- **Step 1:**

  Fit the GLM for the response $Y_1$ on *all* the covariates including $X_1$, $X_2$, and $X_3$ using the IRLS algorithm to obtain the estimates $\widehat{\beta}_{10}^*$, $\widehat{\beta}_{1x_1}^*$, $\widehat{\beta}_{1x_2}^*$, and $\widehat{\beta}_{1x_3}^*$ .

- **Step 2:**

  In **Step 1**, the *pseudo*-response variable $Z_1^*$ (defined below) on the covariates $X_1$, $X_2$,

and $X_3$ on the convergence:

$$\begin{aligned} Z_1^* &= \left(\widehat{\beta}_{10}^* + \widehat{\beta}_{1x_1}^* X_1 + \widehat{\beta}_{1x_2}^* X_2 + \widehat{\beta}_{1x_3}^* X_3\right) + g_1'\left(\widehat{\mu}_1^*\right)\left(Y_1 - \widehat{\mu}_1^*\right) \\ &= g_1\left(\widehat{\mu}_1^*\right) + g_1'\left(\widehat{\mu}_1^*\right)\left(Y_1 - \widehat{\mu}_1^*\right) \quad \left(g_1'\left(\widehat{\mu}_1^*\right)\left(Y_1 - \widehat{\mu}_1^*\right) = \epsilon_{1,2}^*\right) \\ Y_1 &= \widehat{\mu}_1^* + \frac{Z_1^* - g_1\left(\widehat{\mu}_1^*\right)}{g_1'\left(\widehat{\mu}_1^*\right)} \end{aligned}$$

where

$$\widehat{\mu}_1^* = \widehat{Y}_1^* = g_1^{-1}\left(\widehat{\beta}_{10}^* + \widehat{\beta}_{1x_1}^* X_1 + \widehat{\beta}_{1x_2}^* X_2 + \widehat{\beta}_{1x_3}^* X_3\right).$$

- **Step 3:**

Notice that by plugging

$$Y_1 = \widehat{Y}_1^* + \frac{Z_1^* - g_1\left(\widehat{\mu}_1^*\right)}{g_1'\left(\widehat{\mu}_1^*\right)}$$

from **Step 2** into the second equation, we have

$$\begin{aligned} Y_2 &= \beta_{20} + \beta_{2x_1} X_1 + \beta_{2x_3} X_3 + \gamma_{2y_1} Y_1 + \epsilon_2 \\ &= \beta_{20} + \beta_{2x_1} X_1 + \beta_{2x_3} X_3 + \gamma_{2y_1} \widehat{Y}_1^* + \left[\gamma_{2y_1}\left(\frac{Z_1^* - g_1(\widehat{\mu}_1^*)}{g_1'(\widehat{\mu}_1^*)}\right) + \epsilon_2\right] \\ &= \beta_{20} + \beta_{2x_1} X_1 + \beta_{2x_3} X_3 + \gamma_{2y_1} \widehat{Y}_1^* + \epsilon_{2,2}^* \end{aligned}$$

Finally, find the OLS estimator

$$\widehat{\boldsymbol{\beta}}_{2,2SLS} = \left(\mathbf{X}_2^{*\mathrm{T}} \mathbf{X}_2^*\right)^{-1} \mathbf{X}_2^{*\mathrm{T}} \mathbf{Y}_2$$

where $\mathbf{X}_2^* = \left[\mathbf{1}, \mathbf{X}_1, \mathbf{X}_3, \widehat{\mathbf{Y}}_1^*\right]$.

## 3.3 Instrumental Variable (IV) Estimator

- **Steps 1 and 2:**

  These two steps are the same as **Steps 1** and **2** of the 2SLS estimator.

- **Step 3:**

  From the second equation, we have

$$Y_2 = \beta_{20} + \beta_{2x_1} X_1 + \beta_{2x_3} X_3 + \gamma_{2y_1} Y_1 + \epsilon_2$$

  or in matrix notation

$$\mathbf{Y}_2 = \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2.$$

  The instrumental variable (IV) estimator is

$$\widehat{\boldsymbol{\beta}}_{2,IV} = (\mathbf{X}_2^{*\mathrm{T}} \mathbf{X}_2)^{-1} \mathbf{X}_2^{*\mathrm{T}} \mathbf{Y}_2 \tag{4.2}$$

  where $\mathbf{X}_2 = [\mathbf{1}, \mathbf{X}_1, \mathbf{X}_3, \mathbf{Y}_1]$ and $\mathbf{X}_2^* = \left[\mathbf{1}, \mathbf{X}_1, \mathbf{X}_3, \widehat{\mathbf{Y}}_1^*\right]$, $\mathbf{X}_2^*$ are satisfied the following requirements of an IV for estimating $\boldsymbol{\beta}_2$:

$$\begin{aligned}
\operatorname{plim} \frac{1}{n} \mathbf{X}_2^{*\mathrm{T}} \boldsymbol{\epsilon}_2 &= \mathbf{0}, \\
\operatorname{plim} \frac{1}{n} \mathbf{X}_2^{*\mathrm{T}} \mathbf{X}_2 &= \boldsymbol{\Sigma}_{X_2^* X_2} \quad \text{(a finite nonsingular matrix)}, \\
\operatorname{plim} \frac{1}{n} \mathbf{X}_2^{*\mathrm{T}} \mathbf{X}_2^* &= \boldsymbol{\Sigma}_{X_2^* X_2^*} \quad \text{(a positive definite matrix)}.
\end{aligned}$$

Taking $\mathbf{X}_2^*$ as the IV, we can see intuitively that from equation (4.2)

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{2,IV} &= (\mathbf{X}_2^{*\mathrm{T}}\mathbf{X}_2)^{-1}\mathbf{X}_2^{*\mathrm{T}}\mathbf{Y}_2 \\
&= (\mathbf{X}_2^{*\mathrm{T}}\mathbf{X}_2)^{-1}\mathbf{X}_2^{*\mathrm{T}}(\mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2) \\
&= (\mathbf{X}_2^{*\mathrm{T}}\mathbf{X}_2)^{-1}\mathbf{X}_2^{*\mathrm{T}}\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}_2^{*\mathrm{T}}\mathbf{X}_2)^{-1}\mathbf{X}_2^{*\mathrm{T}}\boldsymbol{\epsilon}_2) \\
&= \boldsymbol{\beta}_2 + (\mathbf{X}_2^{*\mathrm{T}}\mathbf{X}_2)^{-1}\mathbf{X}_2^{*\mathrm{T}}\boldsymbol{\epsilon}_2
\end{aligned}$$

is a reasonable estimator of $\boldsymbol{\beta}_2$ as long as $(\mathbf{X}_2^{*\mathrm{T}}\mathbf{X}_2)^{-1}\mathbf{X}_2^{*\mathrm{T}}\boldsymbol{\epsilon}_2$ is somehow close to zero.

We have done the simulations and the results are very promising. Two formal papers are in preparation for publication.