Path diagrams as a notational formalism (Noodling around with pictures)

Vince Wiggins
Vice President, Scientific Development
StataCorp LP

2013 London Stata Users Group Meeting

Introduction

Sometimes path diagrams are used in lieu of mathematical notation to define and explain models.

How far can they be pushed?

Maths and Paths — Covariates

Mathematical notation for a covariate

Maths and Paths — Covariates

Mathematical notation for a covariate

$$x$$

Maths and Paths — Covariates

Mathematical notation for a covariate

$$x$$

Path notation for a covariate

Maths and Paths — Covariates

Mathematical notation for a covariate

$$x$$

Path notation for a covariate

$$\boxed{\mathrm{x}}$$

Maths and Paths — Dependent variables

Mathematical notation for a dependent variable

Mathematical notation for a dependent variable

$$y$$

Maths and Paths — Dependent variables

Mathematical notation for a dependent variable

$$y$$

Path notation for a dependent variable

Maths and Paths — Dependent variables

Mathematical notation for a dependent variable

$$y$$

Path notation for a dependent variable

$$\boxed{\mathrm{y}}$$

Maths and Paths — Parameters/Coefficients

Mathematical notation for a parameter

Maths and Paths — Parameters/Coefficients

Mathematical notation for a parameter

$$\beta$$

Maths and Paths — Parameters/Coefficients

Mathematical notation for a parameter

$$\beta$$

Path notation for a parameter

Maths and Paths — Parameters/Coefficients

Mathematical notation for a parameter

$$\beta$$

Path notation for a parameter

$$\xrightarrow{\quad\quad\beta\quad\quad}$$

Maths and Paths — Parameters/Coefficients

Mathematical notation for a parameter

$$\beta$$

Path notation for a parameter

$$\xrightarrow{\qquad\quad\beta\qquad\quad}$$

Paths are truly parameters in a linear form, and must connect observed or latent variables. So better is

Maths and Paths — Parameters/Coefficients

Mathematical notation for a parameter

$$\beta$$

Path notation for a parameter

$$\xrightarrow{\quad\beta\quad}$$

Paths are truly parameters in a linear form, and must connect observed or latent variables. So better is

$$y = \beta x \ + \ \ldots$$

$$\boxed{x} \xrightarrow{\quad\beta\quad} \boxed{y}$$

Maths and Paths — Latent variables

Mathematical notation for a latent variable

Maths and Paths — Latent variables

Mathematical notation for a latent variable

$$\mu_i \quad \textit{unobserved}$$
$$\mu_i \sim \textit{Normal}(0, \sigma_\mu)$$

Maths and Paths — Latent variables

Mathematical notation for a latent variable

$$\begin{aligned} \mu_i & \quad \textit{unobserved} \\ \mu_i & \sim \textit{Normal}(0, \sigma_\mu) \end{aligned}$$

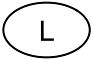
Path notation for a latent variable

Maths and Paths — Latent variables

Mathematical notation for a latent variable

$$\begin{aligned} \mu_i & \quad \textit{unobserved} \\ \mu_i & \sim \textit{Normal}(0, \sigma_\mu) \end{aligned}$$

Path notation for a latent variable



A digression on learning to love latent variables

They can be:

- an error term
- a random intercept
- part of a random slope
- a frailty
- a creator/modeler of endogeneity
- a creator/modeler of sample selection
- a creator/modeler of endogenous treatment effects
- part of a switching model
- or any other latent/unobserved thingy

Linear regression

Mathematical notation for a linear regression

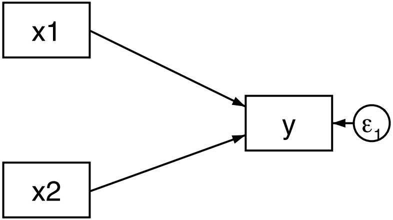
$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \epsilon_i \\
\epsilon_j &\sim \mathit{Normal}(0, \sigma_\epsilon)
\end{aligned}
$$

Linear regression

Mathematical notation for a linear regression

$$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \epsilon_i$$
$$\epsilon_j \sim Normal(0, \sigma_\epsilon)$$

Path notation for a linear regression

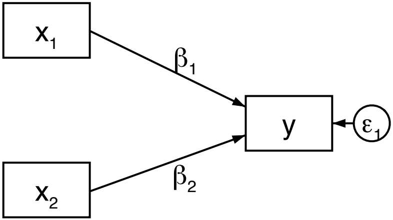


Linear regression

Mathematical notation for a linear regression

$$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \epsilon_i$$
$$\epsilon_j \sim Normal(0, \sigma_\epsilon)$$

Path notation for a linear regression — with coefficients



Linear multilevel model

Mathematical notation for a multilevel model

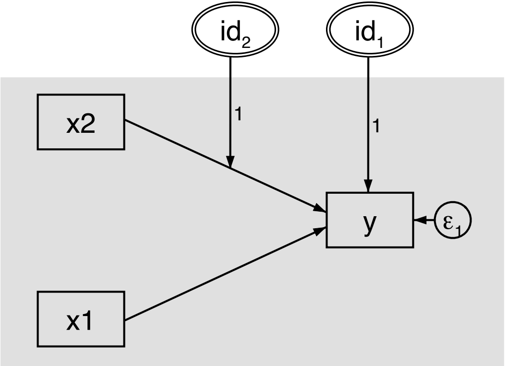
$$
\begin{aligned}
y_{ji} &= \beta_0 + \beta_1 x1_{ji} + (\beta_2 + \mu 1_j) x2_{ji} + \mu 0_j + \epsilon_{ji} \\
\mu 0_j &\sim Normal(0, \sigma_{\mu 0}) \\
\mu 1_j &\sim Normal(0, \sigma_{\mu 1})
\end{aligned}
$$

Linear multilevel model

Mathematical notation for a multilevel model

$$
\begin{aligned}
y_{ji} &= \beta_0 + \beta_1 x1_{ji} + (\beta_2 + \mu 1_j) x2_{ji} + \mu 0_j + \epsilon_{ji} \\
\mu 0_j &\sim \mathit{Normal}(0, \sigma_{\mu 0}) \\
\mu 1_j &\sim \mathit{Normal}(0, \sigma_{\mu 1})
\end{aligned}
$$

Path notation for a multilevel model

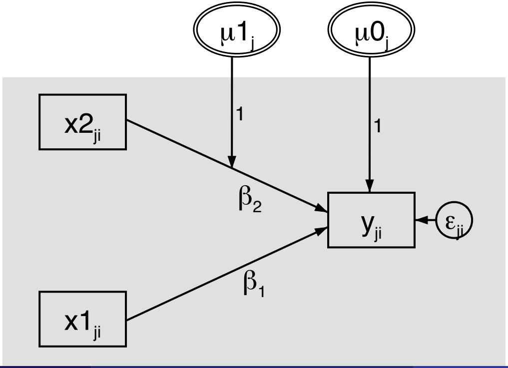


Linear multilevel model

Mathematical notation for a multilevel model

$$\begin{aligned} y_{ji} &= \beta_0 + \beta_1 x1_{ji} + (\beta_2 + \mu1_j)x2_{ji} + \mu0_j + \epsilon_{ji} \\ \mu0_j &\sim Normal(0, \sigma_{\mu0}) \\ \mu1_j &\sim Normal(0, \sigma_{\mu1}) \end{aligned}$$

Path notation for a multilevel model — with maths



A digression on the SEM Builder

Every diagram in this talk was drawn with it.

They each took 2 or 3 minutes to draw.

They can all be estimated by **`sem`** or **`gsem`**.

Let's draw one ...

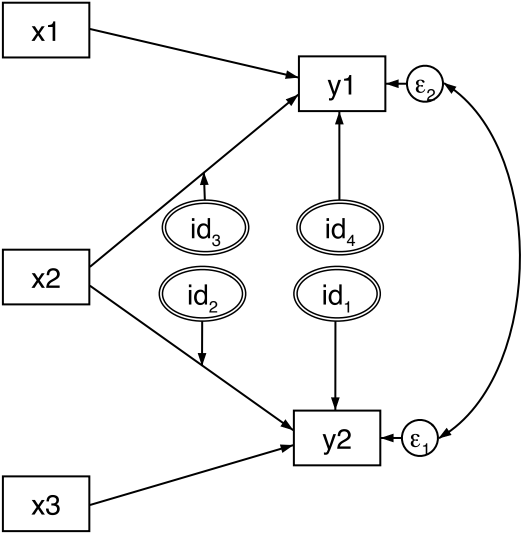
If you are reading this online, you just had to be there.

Linear multivariate (seemingly unrelated) multilevel model

Mathematical notation

$$
\begin{aligned}
y1_{ji} &= \beta_{0a} + \beta_1 x1_{ji} + (\beta_2 + \mu 1_j)x2_{ji} + \mu 0_j + \epsilon 1_{ji} \\
y2_{ji} &= \beta_{0b} + \beta_3 x3_{ji} + (\beta_4 + \mu 2_j)x2_{ji} + \mu 3_j + \epsilon 2_{ji} \\
\mu 0_j &\sim \textit{Normal}(0, \sigma_{\mu 0}) \\
\mu 1_j &\sim \textit{Normal}(0, \sigma_{\mu 1}) \\
\mu 2_j &\sim \textit{Normal}(0, \sigma_{\mu 2}) \\
\mu 3_j &\sim \textit{Normal}(0, \sigma_{\mu 3}) \\
j &\quad \textit{indexes levels}
\end{aligned}
$$

Linear multivariate (seemingly unrelated) multilevel model



Variations on the linear multivariate multilevel model

What if the random effects are shared?

- Easy enough in the mathematical notation, but easier to see in the path diagram. Let's edit the path diagram ...
- *I am sorry once again to those following at home.*

Variations on the linear multivariate multilevel model

What if the random effects are shared?

- Easy enough in the mathematical notation, but easier to see in the path diagram. Let's edit the path diagram ...
- *I am sorry once again to those following at home.*

What if the random coefficients on $x2$ are the same for both y's?

- Again, very easy to see (and to change) on the path diagram ...
- Must explicitly constrain paths/coefficients from latent variables.

Variations on the linear multivariate multilevel model

What if the random effects are shared?

- Easy enough in the mathematical notation, but easier to see in the path diagram. Let's edit the path diagram ...
- *I am sorry once again to those following at home.*

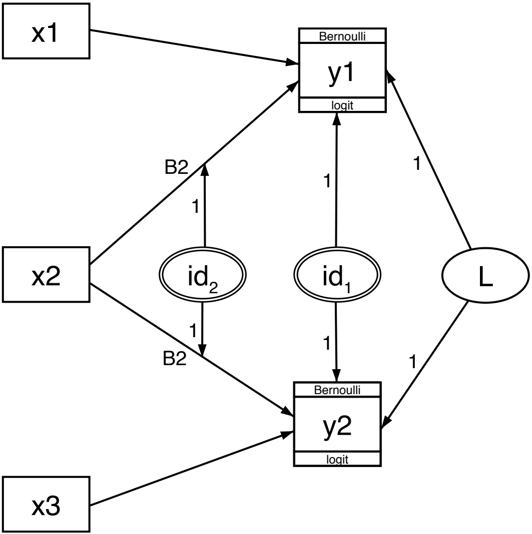
What if the random coefficients on *x*2 are the same for both *y*'s?

- Again, very easy to see (and to change) on the path diagram ...
- Must explicitly constrain paths/coefficients from latent variables.

What if the outcomes were 0/1 and we wanted to model them as logistic?

- We would add some probabilistic statements to our mathematical notation. And likely some verbal explanation.
- We can also designate this on the path diagram ...

Final Variation



Time series — ARMAX model

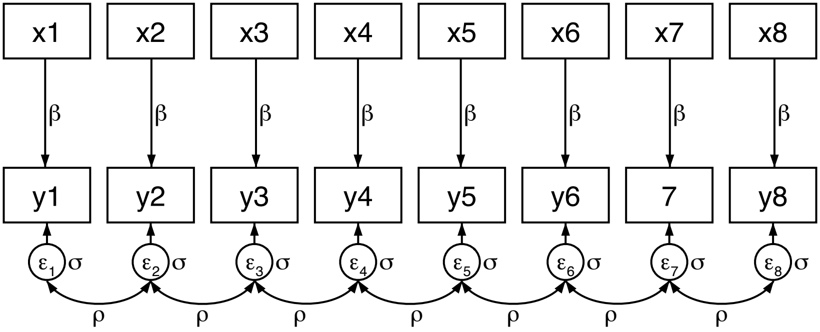
Mathematical notation

$$
\begin{aligned}
y_t &= \beta x_t + \epsilon_t \\
\epsilon_t &= \rho \epsilon_{t-1} + \mu_t \\
\mu_t &\sim Normal(0, \sigma_\mu)
\end{aligned}
$$

Mathematical notation

$$y_t = \beta x_t + \epsilon_t$$
$$\epsilon_t = \rho\epsilon_{t-1} + \mu_t$$
$$\mu_t \sim Normal(0, \sigma_\mu)$$

Path notation



Regression with sample selection — Heckman style

Mathematical notation

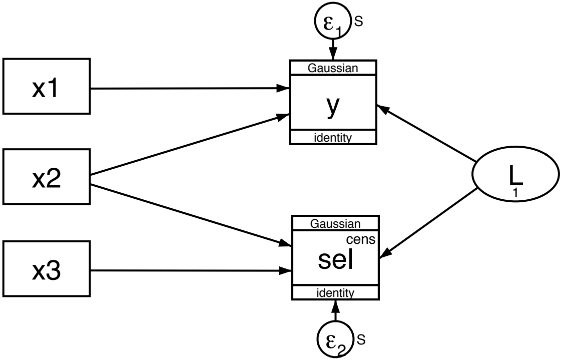
$$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \epsilon_i$$
$$s_i = \lambda_0 + \lambda_2 x2_i + \lambda_3 x3_i + \xi_i$$
$$cov(\epsilon, \xi) \neq 0$$
$$y_i \textit{ observed when } s_i > 0$$

Regression with sample selection — Heckman style

Mathematical notation

$$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \epsilon_i$$
$$s_i = \lambda_0 + \lambda_2 x2_i + \lambda_3 x3_i + \xi_i$$
$$cov(\epsilon, \xi) \neq 0$$
$$y_i \textit{ observed when } s_i > 0$$

Path notation



Logistic model with endogenous covariate

Mathematical notation

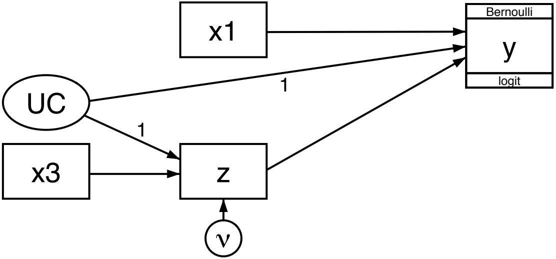
$$Pr(y_i) = logit(\beta_0 + \beta_1 x1_i + \beta_2 z_i + UC_i)$$
$$z_i = \gamma_0 + \gamma_1 x2_i + UC_i + \nu_i$$
$$cov(\epsilon, \xi) = 0$$

Logistic model with endogenous covariate

Mathematical notation

$$Pr(y_i) = logit(\beta_0 + \beta_1 x1_i + \beta_2 z_i + UC_i)$$
$$z_i = \gamma_0 + \gamma_1 x2_i + UC_i + \nu_i$$
$$cov(\epsilon, \xi) = 0$$

Path notation



Selection with an endogenous covariate in a multilevel framework

Regression component

$$y_{ji} = \beta_0 + (\beta_1 + \mu 1_j) x1_{ji} + \beta_2 z_{ji} + UC_{ji} + \mu 0_j + \epsilon_{ji}$$

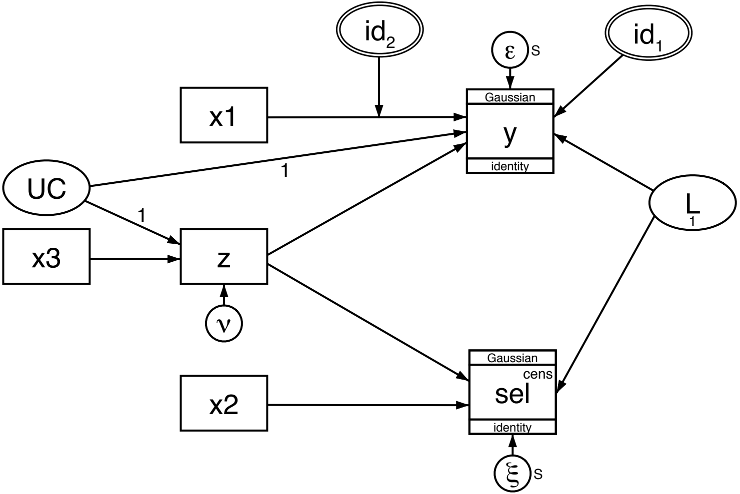
Selection component

$$s_{ji} = \lambda_0 + \lambda_1 z_i + \lambda_2 x2_i + \xi_i$$
$$cov(\epsilon, \xi) \neq 0$$
$$y_{ji} \; observed \; when \; s_{ji} > 0$$

Endogenous component

$$z_j i = \gamma_0 + \gamma_1 x3_j i + UC_j i + \nu_i$$
$$cov(\epsilon, \nu) = 0$$
$$cov(\xi, \nu) = 0$$

Selection with an endogenous covariate in a multilevel framework — path diagram



Closing

Compared to math, path diagrams are:

- easier to understand
- more attuned to human perception
- less precise
- less flexible
- more amenable to noodling

Don't choose.
Use what works — words, maths, paths.