3.26pt

Extending the 3-step approach to a multilevel SEM

Yajing Zhu

Department of Statistics
London School of Economics and Political Science, UK

10 July 2018

Substantive research question

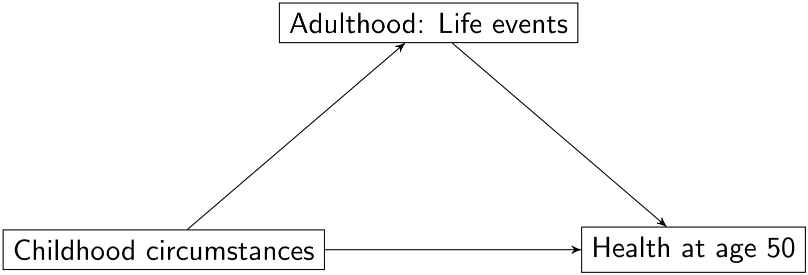

Figure 1: A general joint modelling framework to explore the potential pathways between childhood circumstances, life events and health in mid-life.

Description of the dataset: recently published sweep NCDS9 (2013-2014, age 55) achieved 9,125 CMs

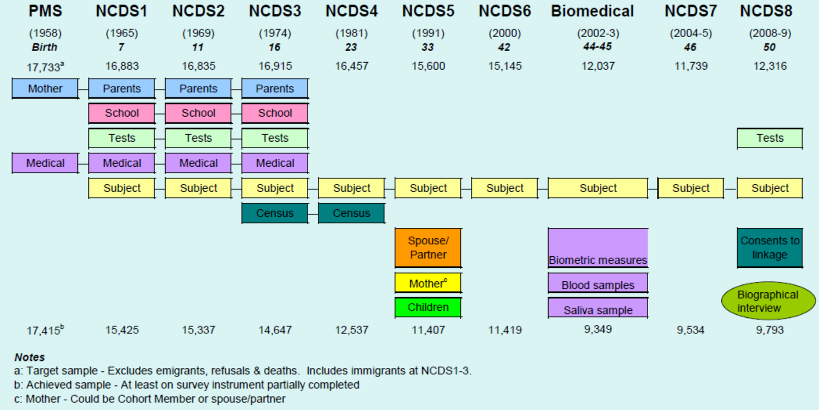

Figure 2: Overview of the dataset

Methodological challenges

- Multiple repeated measures of several aspects of childhood socio-economic circumstances (SECs) at ages 0, 7, 11 and 16
  - Latent class models to characterise the patterns of change in each dimension of childhood SECs
- Relate multiple and possibly associated latent categorical variables to temporally distal outcomes of mixed types and measured at different levels (life events, midlife health)
- Misclassification error in the latent class model, endogeneity, missing data
  - 3-step approach, multilevel structural equation model

Review of previous work

Main interest

How to include latent summaries of childhood SECs as predictors of a distal outcome?

- 1-step approach
  - Problem: unintended circular relationship.
- naive 3-step approaches (modal class, pseudo class)
  - Problem: misclassification, underestimated/overestimated standard errors.
- Advanced 3-step approaches (modified BCH, Lanza's approach, ML)

A general 3-step ML approach I

- 1-LV: Vermunt (2010), Asparouhov and Muthén (2014), Bakk and Vermunt (2016)
- Multiple LVs: Zhu et al. (2017): generalisation& robustness test.

Steps

- Step 1: Estimate separate latent class models for categorical predictors.
- Step 2: Calculate misclassification probabilities.
- Step 3: Estimate models of interest, with categorical LVs as predictors.

A general 3-step ML approach II

- Notation: Cs for childhood SECs; Ys for indicators for Cs; Ms for the most likely class membership; H for the distal outcome.
- Assumption: $C_1 \perp\!\!\!\perp M_2 | C_2$; $C_2 \perp\!\!\!\perp M_1 | C_1$, $Z \perp\!\!\!\perp Ys | C_1, C_2$.

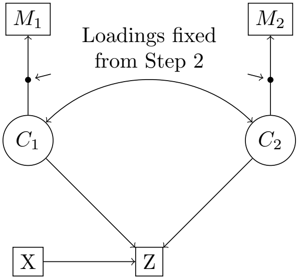

Figure 3: The 3-step approach with two latent categorical variables C_1 and C_2.

Childhood SECs $\rightarrow$ time-to-event outcomes I

Discrete-time survival data (partnership transitions)

- Denote by y_{ij} the duration of episode j of individual i, which is fully observed if an event occurs ($\delta_{ij} = 1$) and right-censored if not ($\delta_{ij} = 0$).
- Data restructuring: convert the observed data (y_{ij}, δ_{ij}) to a sequence of binary responses (y_{tij}), indicating whether an event has occurred in time interval $[t, t + 1)$.
- Discrete-time hazard function: $h_{tij} = Pr(y_{tij} = 1 | y_{t' < t, ij} = 0)$.

Childhood SECs $\rightarrow$ time-to-event outcomes II

Step 3 is a random effects logit model, allowing for a log-linear structure between LVs.

$$\log\left(\frac{h_{tij}}{1-h_{tij}}\right) = \alpha_t + \boldsymbol{\beta}'\mathbf{X}_{tij} + \sum_{q=1}^{Q}\sum_{k_q=1}^{K_q-1}\tau_{C_q,k_q} I(C_{qi}=k_q) + u_i$$

- h_{tij} is the hazard of partnership transitions (formation and dissolutions)
- α_t is the baseline hazard function
- $\mathbf{X}_{tij}$ is the vector of time-varying and time-invariant predictors
- τ_{C_q,k_q}s are the class-specific coefficients of LV C_q
- $u_i \sim N(0,\sigma_u^2)$ is the individual-specific unobserved random effect

Structural equation models

Recall the substantive research question:

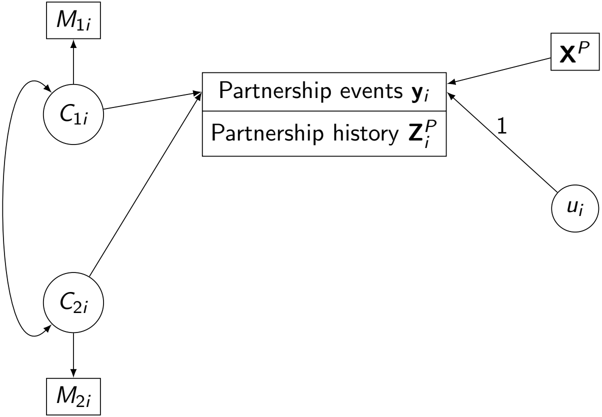

Figure 4: A general path diagram of a multilevel SEM with factorised individual-level random effects.

Structural equation models

Recall the substantive research question:

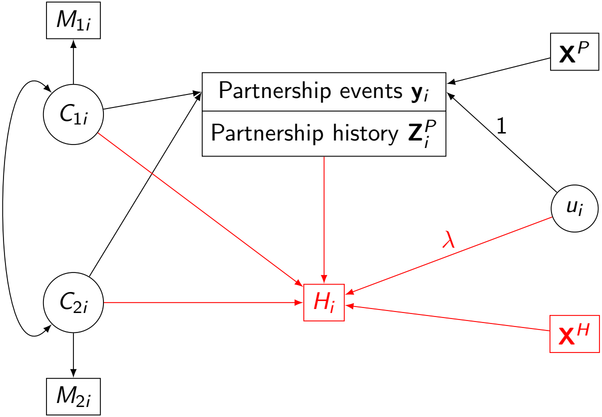

Figure 4: A general path diagram of a multilevel SEM with factorised individual-level random effects.

SEM: Model specification

Joint modelling of multilevel outcomes of mixed types that are predicted by latent categorical variables.

$$\text{logit}\Big(h_{tij}\Big) = \alpha_t + \sum_{k_q=1}^{K_q-1} \alpha_{C_q,k_q} I(C_{qi} = k_q) + \boldsymbol{\alpha}'\mathbf{X}_{tij}^{(P)} + u_i,$$

$$\text{logit}\Big(P(H_i = 1)\Big) = \beta_0 + \sum_{k_q=1}^{K_q-1} \beta_{C_q,k_q} I(C_{qi} = k_q) + \boldsymbol{\beta}_1'\mathbf{X}_i^{(H)} + \boldsymbol{\beta}_2'\mathbf{Z}_i^{(P)} + \lambda u_i.$$

- H_i is binary health status, 1=poor health.
- $\mathbf{X}_i^{(H)}$ is a vector of health-relevant covariates.
- $\mathbf{X}_{tij}^{(P)}$ is a vector of predictors of separation hazard.
- $\mathbf{Z}_i^{(P)}$ is a vector of summary indicators of partnership stability derived from the partnership history (e.g. # partners during ages 16-50, % time single).

Advantages of the framework

- The 3-step approach handles the misclassification error in the latent class model.
- Joint modelling handles endogeneity of $\mathbf{Z}_i^{(P)}$ in the health model.
- Allow for differential effects (λ) of a common set of individual-specific unobservables (u_i) on the hazard of separation and health.
- u_i also accounts for the additional dependence between outcomes that are not accounted for by covariates and latent class variables.
- Residual correlations can be computed from λs.
- Generalisability: data with complex structures (e.g. multilevel, longitudinal, mixed response types), dropout mechanism and related processes, multiple health outcomes $\Rightarrow$ better identification of σ_u^2.

Simulation results I

Data generating model:

- Generate 2-category latent classes C_1 and C_2 from the log-linear model and the manifest variable Ys (tweak $Y|C$ parameters for different levels of class separation)
- Generate discrete time-to-event data for repeatable events.
- Generate a single binary health outcome that is predicted by a summary of life events (i.e. total # of events) and Cs.

Models are estimated in LatentGOLD 5.1 for settings with combinations of sample sizes (N=500, 2000) and entropy values (0.8 and 0.4) of the measurement models.

Simulation results II

Findings:

- Relative bias in estimates is less than 1% and all less than 5% for the scenario with good classification (high entropy) and large sample size ($N = 2000$). Average standard errors are very close to the standard deviations and the nominal coverage for all parameters is close to the expected level of 95%.
- Small sample size ($N = 500$) and low entropy (0.4): estimates of the coefficients has a relative bias above 10% $\rightarrow$ scaling effect due to the the biased estimate for σ_u^2 (17.8% relative bias) as the magnitude of coefficients depends on the magnitude of random effect variance in a random effect GLM.
- Increase the N from 500 to 2000, despite having poor classification in the measurement models, both the point estimates and standard errors improve.
- Across all scenarios investigated, estimates in the event history submodel are in general less biased than those in the health submodel: multiple observations per individual for event history outcome while health outcome observed only once for an individual.

Simulation results III

- Implication 1: To estimate an SEM with such a complex structure, with multiple individual-level latent categorical variables and one individual-level latent continuous variable, large N & a large number of lower level units, are necessary for model identification. More health outcomes (i.e. more individual-level indicators) or repeated measures of health outcomes (i.e. more time-varying indicators) could both be beneficial.
- Implication 2: A large number of latent class variables can lead to heavy computational time (integration) and loss in precision due to numerical approximation schemes.

Substantive findings

Model 1 is the health model alone, Model 2 is the generalised SEM with a submodel for the time to dropout.

Health submoodel	Model 1		Model 2	
Covariates	Est.	(SE)	Est.	(SE)
Intercept	-2.36**	(0.09)	-3.06**	(0.23)
Overweight[1] (ref.= No)	0.25**	(0.07)	0.26**	(0.07)
Childhood circumstances				
Social class[2] (ref.=High)				
Low	0.40**	(0.19)	0.44**	(0.12)
Medium	0.32**	(0.11)	0.30**	(0.10)
Financial difficulty (ref.=Low)				
High	0.53**	(0.21)	0.42**	(0.10)
Material hardship (ref.=Low)				
Medium	0.33**	(0.11)	0.32**	(0.09)
High	0.35**	(0.12)	0.39**	(0.10)
Family structure (ref.=Stable)				
Unstable	0.08	(0.13)	0.17	(0.17)
Partnership experience				
Total number of partners before age 50 (ref. =1)				
0			-0.13	(0.32)
2			0.18	(0.14)
3+			0.41*	(0.24)
Age at first partnership			-0.13**	(0.05)
Percentage time spent single			1.26**	(0.38)
Random effect parameters				
σ_u^2			1.32**	(0.10)
$\lambda^{(H)}$			-0.44**	(0.16)
$\lambda^{(F)}$			-0.05**	(0.02)
$\lambda^{(D)}$			1.25**	(0.12)

** $p < 0.05$, * $p < 0.1$

[1] Binary indicator for overweight at age 16.

[2] Father or male head social class.

References

Asparouhov, T. and Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(3):329–341.

Bakk, Z. and Vermunt, J. K. (2016). Robustness of stepwise latent class modeling with continuous distal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(1):20–31.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*, volume 904. John Wiley & Sons.

Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):667–678.

Vermunt, J. K. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18(4):450–469.

Zhu, Y., Steele, F., and Moustaki, I. (2017). A general 3-step maximum likelihood approach to estimate the effects of multiple latent categorical variables on a distal outcome. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(5):643–656.

Appendix: More on estimation I

1) Estimation for ω parameters in the log-linear model

Denote by parameter vector $\theta_1 = (\omega_0, \omega_{k_1}^{C_1}, \omega_{k_2}^{C_2}, \omega_{k_1 k_2}^{C_1 C_2})$. The individual contribution to the expected score function of θ_1 can be written:

$$E\left[S_i(\theta_1)\right] = \sum_{k_2=1}^{K_2} \sum_{k_1=1}^{K_1} \int_u S_i(\theta_1) g(\xi_i | \mathbf{y}_{tij}, \mathbf{X}_{tij}, M_{1i}, M_{2i}) du. \tag{1}$$

In the M-step, we need to solve $\sum_{i=1}^{N} E\left[S_i(\theta_1)\right] = 0$. Integrals in (1) can be approximated by, e.g. Monte Carlo methods (Sammel et al., 1997) or Gaussian-Hermite quadratures which replaces the integral with a weighted summation over u_i.

Appendix: More on estimation II

2) Estimation for parameters in the survival model

Denote by parameter vector $\theta_2 = (\alpha_t, \beta, \tau_{k_1}, \tau_{k_2}, \sigma_u)$, the individual contribution to the expected score function of θ_2 is

$$E\left[S_i(\theta_2)\right] = \sum_{k_2=1}^{K_2-1} \sum_{k_1=1}^{K_1-1} \int_u S_i(\theta_2) g(\xi_i | \mathbf{y}_{tij}, \mathbf{X}_{tij}, M_{1i}, M_{2i}) du. \qquad (2)$$

Similar to earlier practices, solving $\sum_{i=1}^{N} E\left[S_i(\theta_2)\right] = 0$ requires the approximation of the integral in (2). Higher dimensions of the latent variables (either discrete or continuous) can be computationally expensive.

Appendix: More on estimation III

Summary of estimation in Step 3:

1. Generate initial estimates for all parameters (θ_1, θ_2).
2. E-step: compute $E\left[S_i(\theta_1)\right]$ and $E\left[S_i(\theta_2)\right]$ given in (1) and (2).
3. M-step: solve for $\sum_{i=1}^{N} E\left[S_i(\theta_1)\right] = 0$ and $\sum_{i=1}^{N} E\left[S_i(\theta_2)\right] = 0$, update parameter estimates.
4. Repeat steps 2 and 3 until convergence is reached.

Standard errors:
Denote by vector $\theta = (\theta_1, \theta_2)$. To obtain asymptotic standard errors: compute the information matrix $I(\theta)$ using maximum likelihood estimates; take diagonal elements of the inverse of $I(\hat{\theta})$. An alternative: use parametric bootstrap methods that are available in many software packages (Bartholomew et al., 2011).

Aeppendix: More on simulation results I

Table 1: Simulation results for the 3-step procedure applied to joint model for a event history submodel and a distal health submodel: high entropy (0.8) and small sample size ($N = 500$)

Parameter	True	Relative bias (%)	SE	SD	95% Coverage
		Event history submodel			
β_0	-2.00	0.34	0.23	0.22	0.96
$\beta_1(t)$	1.50	0.03	0.11	0.11	0.96
$\beta_2(X^{(P)})$	1.50	0.11	0.10	0.10	0.95
$\beta_3(X_t^{(P)})$	-0.50	0.28	0.06	0.06	0.96
$\tau_1^{C_1}$	2.50	0.27	0.18	0.18	0.95
$\tau_1^{C_2}$	-1.00	-1.07	0.17	0.18	0.95
		Health submodel			
α_0	-3.00	5.49	1.03	2.15	0.86
$\alpha_1(X^{(H)})$	-1.50	6.19	0.50	1.14	0.83
$\alpha_2(Z^{(P)})$	0.50	6.11	0.17	0.35	0.86
$\gamma_1^{C_1}$	-2.00	4.15	0.91	1.55	0.88
$\gamma_1^{C_2}$	1.50	2.24	0.81	1.39	0.90
λ	1.50	0.55	1.95	2.33	0.71
σ_u^2	1.00	-1.87	0.22	0.21	0.94
ω_{12}	-0.50	-0.46	0.23	0.22	0.97

Relative bias (%)= (Estimate-True) / True$\times$100%

Aeppendix: More on simulation results II

Table 2: Simulation results for the 3-step procedure applied to joint model for a event history submodel and a distal health submodel: high entropy (0.8) and large sample size ($N = 2000$)

Parameter	True	Relative bias (%)	SE	SD	95% Coverage
		Event history submodel			
β_0	-2.00	0.25	0.12	0.12	0.95
$\beta_1(t)$	1.50	0.26	0.06	0.05	0.96
$\beta_2(X^{(P)})$	1.50	0.07	0.05	0.05	0.96
$\beta_3(X_t^{(P)})$	-0.50	0.25	0.03	0.03	0.96
$\tau_1^{C_1}$	2.50	0.01	0.09	0.09	0.95
$\tau_1^{C_2}$	-1.00	-0.19	0.09	0.09	0.95
		Health submodel			
α_0	-3.00	1.42	0.30	0.30	0.96
$\alpha_1(X^{(H)})$	-1.50	1.89	0.15	0.15	0.96
$\alpha_2(Z^{(P)})$	0.50	1.02	0.05	0.05	0.97
$\gamma_1^{C_1}$	-2.00	1.34	0.25	0.24	0.97
$\gamma_1^{C_2}$	1.50	1.64	0.22	0.22	0.96
λ	1.50	3.10	0.29	0.29	0.96
σ_u^2	1.00	0.06	0.11	0.11	0.94
ω_{12}	-0.50	-1.24	0.11	0.12	0.93

Relative bias (%)= (Estimate-True) / True$\times$100%

Aeppendix: More on simulation results III

Table 3: Simulation results for the 3-step procedure applied to joint model for a event history submodel and a distal health submodel: low entropy (0.4) and small sample size ($N = 500$)

Parameter	True	Relative bias (%)	SE	SD	95% Coverage
		Event history submodel			
β_0	-2.00	-1.91	0.31	0.38	0.88
$\beta_1(t)$	1.50	0.24	0.11	0.11	0.96
$\beta_2(X^{(P)})$	1.50	0.33	0.11	0.11	0.95
$\beta_3(X_t^{(P)})$	-0.50	0.03	0.06	0.06	0.96
$\tau_1^{C_1}$	2.50	-3.65	0.25	0.27	0.94
$\tau_1^{C_2}$	-1.00	-2.95	0.31	0.32	0.94
		Health submodel			
α_0	-3.00	11.24	0.81	1.47	0.97
$\alpha_1(X^{(H)})$	-1.50	13.20	0.41	0.95	0.95
$\alpha_2(Z^{(P)})$	0.50	11.16	0.13	0.24	0.98
$\gamma_1^{C_1}$	-2.00	12.49	0.66	1.14	0.98
$\gamma_1^{C_2}$	1.50	12.86	0.57	0.89	0.97
λ	1.50	22.51	0.81	1.60	0.95
σ_u^2	1.00	17.83	0.32	0.40	0.89
ω_{12}	-0.50	-11.84	0.52	0.51	0.97

Relative bias (%)= (Estimate-True) / True$\times$100%

Aeppendix: More on simulation results IV

Table 4: Simulation results for the 3-step procedure applied to joint model for a event history submodel and a distal health submodel: low entropy (0.4) and large sample size ($N = 2000$)

Parameter	True	Relative bias (%)	SE	SD	95% Coverage
		Event history submodel			
β_0	−2.00	−0.66	0.15	0.17	0.92
$\beta_1(t)$	1.50	0.09	0.06	0.05	0.96
$\beta_2(X^{(P)})$	1.50	−0.07	0.06	0.06	0.94
$\beta_3(X_t^{(P)})$	−0.50	0.30	0.03	0.03	0.95
$\tau_1^{C_1}$	2.50	−0.52	0.12	0.12	0.94
$\tau_1^{C_2}$	−1.00	−0.79	0.15	0.15	0.94
		Health submodel			
α_0	−3.00	0.21	0.43	0.42	0.93
$\alpha_1(X^{(H)})$	−1.50	−0.24	0.20	0.20	0.92
$\alpha_2(Z^{(P)})$	0.50	−0.57	0.07	0.06	0.92
$\gamma_1^{C_1}$	−2.00	−1.37	0.39	0.38	0.93
$\gamma_1^{C_2}$	1.50	0.37	0.36	0.35	0.96
λ	1.50	−3.70	0.41	0.41	0.90
σ_u^2	1.00	3.13	0.15	0.17	0.93
ω_{12}	−0.50	−3.39	0.26	0.26	0.96

Relative bias (%)= (Estimate-True) / True×100%