Application of Statistical Modeling – A very short summary of some key points

Why multilevel modeling:

1. With clustering structures, observations are correlated. The i.i.d assumption of OLS is violated.
2. Effective sample size of clustering data is smaller. $SE = \frac{\sigma^2}{\sqrt{n}}$ ↑ under over more likely to have significant results. multilevel analyses correct for this.

multilevel, model specification

Empty model:

$$y_{ij} = \beta_{0j} + \varepsilon_{ij} \quad \text{level 1} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\beta_{0j} = \beta_0 + u_{0j} \quad \text{level 2} \quad u_{0j} \sim N(0, \sigma_0^2)$$

One covariate:

$$y_{ij} = \beta_{0j} + X_{1i}\beta_{1j} + \varepsilon_{ij} \quad \text{level 1} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\left.\begin{aligned}\beta_{0j} &= \beta_0 + u_{0j} \\ \beta_{1j} &= \beta_1 + u_{1j}\end{aligned}\right\} \text{level 2} \qquad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} = D \right)$$

mixed model:

$$y_{ij} = X_r\beta_{pj} + X_f\beta_k + \varepsilon_{ij} \quad \text{level 1}$$

where X_r are variables that we specify random effects for

X_f are variables that we specify fixed effect for

$$\beta_{pj} = \beta_p + u_j$$

$$u_j = \begin{bmatrix} u_{0j} \\ u_{1j} \\ \vdots \\ u_{pj} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \cdots & \sigma_{0p} \\ \sigma_{10} & \sigma_1^2 & \vdots \\ & \ddots & \vdots \\ \sigma_{p0} & & \sigma_p^2 \end{bmatrix} \right)$$

β_k are parameters to be estimated

$\varepsilon_{ij} \sim N(0, \sigma^2)$

$$y_{ij} \mid u_j \sim N\left(X_{ij}\beta + X_{ri}u_j \;, \sigma^2 \right)$$

$$y_j \mid u_j \sim N_{nj}\left(X_j\beta + X_{rj}u_j \;, R_j \right)$$

$$y_{ij} = X_r \beta_{pj} + X_f \beta_k + \varepsilon_{ij}$$

where X_r are (intercept + covariates) with random effects
X_f are covariates with fixed effects
ε_{ij} is the random error

Conditional specification

$$y_j \mid u_j \sim N(X_j \beta + X_{rj} u_j, R_j)$$

① X_j is the matrix of covariates

n_j cluster size

$$\begin{bmatrix} X_{11} & X_{21} & \cdots & X_{p1} & \cdots & X_{k1} \\ X_{12} & X_{22} & \cdots & X_{p2} & \cdots & X_{k2} \\ \vdots & & & & & \\ X_{1nj} & & & X_{pnj} & & X_{knj} \end{bmatrix}$$

p covariate with random effect

k covariates with fixed effect

② β is a vector of coefficients

$(p+k) \times 1$

$$\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \\ \vdots \\ \beta_k \end{bmatrix}$$

p coefficients as the mean of the random coefficients

k coefficients as fixed effects

③ X_{rj} is part of X_j — the covariate matrix of the covariates that have random effect

n_j cluster size

$$\begin{bmatrix} X_{11} & X_{21} & \cdots & X_{p1} \\ X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & & & \\ X_{1nj} & & & X_{pnj} \end{bmatrix}$$

p covariates with random effect

④ u_j is a vector of the random effect of the cluster

$p \times 1$

$$\begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_p \end{bmatrix}$$

⑤ R_j is the variance covariance matrix of ε_j

$$\varepsilon_j = \begin{bmatrix} \varepsilon_{1j} \\ \varepsilon_{2j} \\ \vdots \\ \varepsilon_{nj\,j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \underbrace{\begin{bmatrix} \sigma^2 & & 0 \\ & \sigma^2 & \\ 0 & & \ddots \\ & & & \sigma^2 \end{bmatrix}}_{R_j} \right)$$

Marginal specification

$$y_j = \begin{bmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{nj\,j} \end{bmatrix} \qquad y_{ij} = X_{ri}\beta_{pj} + X_{fi}\beta_k + \varepsilon_{ij} \qquad \text{where } \beta_{pj} = \beta_p + u_j$$

At the element level

Variance:

$$V(y_{ij}) = V(X_{ri}\beta_{pj} + X_{fi}\beta_k + \varepsilon_{ij})$$
$$= V(X_i\beta + X_{ri}u_j + \varepsilon_{ij}) \quad \text{where } X_i = X_{ri} + X_{fi},\ \beta = \beta_p + \beta_k$$
$$= X_{ri}^2 V(u_j) + V(\varepsilon_{ij})$$

As for covariance:

$y_{1j} = X_{1j}\beta + X_{1rj}u_j + \varepsilon_{1ij}$ — case 1 in cluster j

$y_{2j} = X_{2j}\beta + X_{2rj}u_j + \varepsilon_{1ij}$ — case 2 in cluster j

$$cov(y_{1j}, y_{2j}) = cov(X_{1j}\beta + X_{1rj}u_j + \varepsilon_{1j},\ X_{2j}\beta + X_{2rj}u_j + \varepsilon_{2j})$$
$$= \underbrace{cov(X_{1rj}u_j,\ X_{2rj}u_j)}_{} + \cancel{cov(X_{1rj}u_j,\ \varepsilon_{2j})}^{\,0} + \cancel{cov(\varepsilon_{1j},\ X_{2rj}u_j)}^{\,0} + \cancel{cov(\varepsilon_{1j},\ \varepsilon_{2j})}^{\,0}$$

covariance between two cases in the same cluster is non zero bc they share the r.v. u_j

At the cluster level

$$y_j = X_{rj}\beta_{pj} + X_{fj}\beta_k + \varepsilon_j$$

$$V(y_j) = V(X_{rj}\beta_{pj} + X_{fj}\beta_k + \varepsilon_j)$$
$$= V(X_{rj}\beta_p + X_{rj}u_j + X_{fj}\beta_k + \varepsilon_j) \quad \text{where } u_j \text{ is a } p\times 1 \text{ matrix denoting } p \text{ random variables}$$
$$= V(X_j\beta + X_{rj}u_j + \varepsilon_j)$$
$$= V(X_{rj}u_j + \varepsilon_j)$$
$$= X_{rj}V(u_j)X_{rj}' + V(\varepsilon_j)$$
$$= \underbrace{X_{rj}DX_{rj}' + R_j}_{V_j}$$

marginal variance-covariance matrix.

Multilevel model implies a marginal model.

$V(y_j) = X_{rj} D X_{rj}' + R_j$ where $D = V(u_j)$

the covariance matrix of random effects

As a covariance matrix, D needs to be positive definite

consider the empty model:

$y_{ij} = \beta_{0j} + \epsilon_{ij}$ $\epsilon_{ij} \sim N(0, \sigma^2)$

$\beta_{0j} = \beta_0 + u_{0j}$ $u_{0j} \sim N(0, \sigma_0^2)$

$cov(y_{1j}, y_{2j}) = cov(\beta_0 + u_{0j} + \epsilon_{1j}, \beta_0 + u_{0j} + \epsilon_{2j}) = cov(u_{0j}) = \sigma_0^2 \geq 0$

$corr(y_{1j}, y_{2j}) = \frac{cov(y_{1j}, y_{2j})}{\sqrt{V(y_{1j})}\sqrt{V(y_{2j})}} = \frac{\sigma_0^2}{V(y_{ij})} = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2} \geq 0$

↑

Icc : Intra-class correlation.

must larger than 0 bc the correlation is a function of variance

The off-diagonal elements of $V(y_j)$ matrix are non-negative bc these covariances are a function of the variance.

⇓

$V(y_j)$ is constrained

⇓

multilevel models are less flexible than marginal models

if we directly specify V matrix and do not factor out the u_{0j} part, we are not subjected to this constraint.

It is possible to specify negative Icc.

But in the context of marginal model, we don't have random effect.

we only have correlated variance structure

Pros of multilevel model. D matrix is of interest.

E.g., variance at the higher level.

Pros of marginal model. Relax the constraint on the positive intra-class correlation.

- Estimation uses two-step maximum likelihood:
  1. Estimate V matrix first
  2. with the estimated V $\hat{\beta} = (X'VX)^{-1}X'\hat{V}Y$

- Variance can be computed using the standard GLS results
  $V(\hat{\beta}) = (X'V^{-1}X)^{-1}$
  ⇒ variance of the coefficients is a direct function of V.
  if D and R are specified badly or if V is specified badly, then SE of $\hat{\beta}$ would be bad.

- The estimations use maximum likelihood.
  nested models can be compared with likelihood ratio test.
  Δ -2 log-likelihood $\sim \chi^2_q$ where q is the different # of parameters

Example:
M1: Random intercept model.
M2: Random intercept + one random slope model } difference is σ_1^2 and σ_{01}

$df = 2$?

Not exactly, σ_1^2 is not freely estimated, has to be > 0
∴ more like 1.5 df.

∴ 0.5 (1 - pchisq(test statistics, df = 1)) + 0.5 (1 - pchisq(test statistics, df = 2))
⇒ A mixture p value

- Predicted random effect

Empirically Best Linear Unbiased Predictors (EBLUP) of random effects

The variance and mean of random effects are estimated.
⇒ We can compute their predicted values based on the data.

$$u_j = E(u_j \mid Y_j = y_j) = D Z_j' V_j^{-1} (y_j - X_j \beta)$$

For a demonstration, consider

$$Y_{ij} = \beta_{0j} + X_{1i} \beta_{1j} + \varepsilon_{ij} \qquad \hat{D}$$

$$\beta_{0j} = \beta_0 + u_{0j} \qquad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right)$$

$$\beta_{1j} = \beta_1 + u_{1j}$$

$$Z_j = X_{rj} = \underbrace{\begin{bmatrix} 1 & X_{11} \\ 1 & X_{12} \\ \vdots & \vdots \\ 1 & X_{1 n_j} \end{bmatrix}}_{\text{2 random variables}} \Big\} n_j \qquad R_j = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}$$

$$\hat{u}_j = \begin{bmatrix} \hat{u_{0j}} \\ \hat{u_{1j}} \end{bmatrix} = \underbrace{\underbrace{D}_{2\times 2} \underbrace{Z_j'}_{2\times n_j} \underbrace{V_j^{-1}}_{n_j \times n_j} \underbrace{(y_j - X_j \beta)}_{n_j \times 1}}_{2 \times 1}$$

different clusters'
- residuals are different
- Z_j are different
- V_j are different bc Z_j is included

⇓
get different estimates

- Diagnostics

Residuals of the model $\hat{\varepsilon}_{ij}$ show a normal distribution?

EBLUP computed random effects show a normal distribution?

Conceptually understanding multilevel model as partial pooling:

50 states. intercept only model.

Option 1: Ignore states

$$y_{ij} = \bar{y} + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

i is individuals
j is states.

Option 2: Separate model for the 50 states

$$y_{i1} = \bar{y}_1 + \varepsilon_{i1} \quad \varepsilon_{i1} \sim N(0, \sigma_1^2)$$

$$y_{i2} = \bar{y}_2 + \varepsilon_{i2} \quad \varepsilon_{i2} \sim N(0, \sigma_2^2)$$

$$\vdots$$

$$y_{i50} = \bar{y}_{50} + \varepsilon_{i50} \quad \varepsilon_{i50} \sim N(0, \sigma_{50}^2)$$

Option 3: multilevel

$$y_{ij} = \bar{y}_j + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

$$\bar{y}_j = \bar{y} + u_j \quad u_j \sim N(0, \sigma_u^2)$$

with EBLUP, $\hat{\bar{y}}_j$ can be computed for each state.

This $\hat{\bar{y}}_j$ is a weighted sum of $\bar{y}$ in option 1 and $\bar{y}_j$ in option 2.

Incorporating weights in multilevel models

Model-based approach and including weights as covariates doesn't always make sense.
(assuming that weights carry some information relevant to Y and use them as covariates in order to control for it)

Hybrid approach. need weights at both the levels.

- Higher level selection probability → inverse.
- conditioning on the higher level, lower level selection probability → inverse

E.g., $P(\text{student} \mid \text{school}) = \frac{P(\text{student}, \text{school})}{P(\text{school})}$ ← survey report individual selection weight

weights at the lower level need to be scaled.
sum to effective sample size or
sum to sample size.

Longitudinal time points nested within individuals

If we follow the multilevel approach

$$y_{ti} = \beta_{0i} + \beta_{1i} x_{ti} + \varepsilon_{ti} \qquad \varepsilon_{ti} \sim N(0, \sigma^2) \Rightarrow R_i = \begin{bmatrix} \sigma^2 & & 0 \\ & \sigma^2 & \\ 0 & & \ddots \; \sigma^2 \end{bmatrix}$$

$$\beta_{0i} = \beta_0 + u_{0i}$$

$$\beta_{1i} = \beta_1 + u_{1i} \qquad \begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right)$$

↑ logitudinal analysis typically allows different error structure than this.

options:

1. First-order autoregressive
   - correlations weaken as a power function of time

	T_1	T_2	T_3	T_4
T_1	σ^2	σ_H	σ_M	σ_L
T_2	σ_H	σ^2		
T_3	σ_M		σ^2	
T_4	σ_L			σ^2

2. simple diagonal (constant variance, zero covariance)
3. compound symmetry (constant variance, constant covariance)
4. heterogeneous variance (variance can also change as a function of time)

Marginal models

the correlated variance structure is something to be controlled, but not of research interest.

Researchers are not interested in the variance of the slope, but only the fixed effect applicable <u>across all subjects</u>.

$y_i \sim N_{ni}(X_i\beta, V_i)$ where β is a vector of fixed effect
V_i is a $n_i \times n_i$ matrix

Alternative method for fitting this kind of model:

- marginal linear model (MLM)
- generalized estimating equations (GEE)

Marginal linear model (MLM)

same as multilevel model. a two-step approach

First, estimate V_i matrix

Then, GLS to compute $\hat{\beta} = (X'VX)^{-1}X'VY$

But MLM cannot deal with non-normal outcome variables

GEE is a more flexible alternative.

solve the score function $S(\beta) = \sum^{n} D_i^T V_i^{-1}(y_i - \mu_i) = 0$

where D_i is a $n_i \times p$ matrix with the (i,j)th element being $\frac{\partial \mu_{ti}}{\partial \beta_j}$

μ_i is the expected value based on the specified model

The score function needs a V_i which we do not know.

∴ we work with a "working correlation matrix"

The estimate of β remains robust even with a bad choice of V but standard error of the estimate would be influenced.

Options for the working correlation matrix:

- independence
- exchangeable (constant correlation)
- first-order auto-regressive. decaying correlation over time
- unstructured.

with μ_i being the expected value, the score function can be applied to situations where outcome variables are not normal.

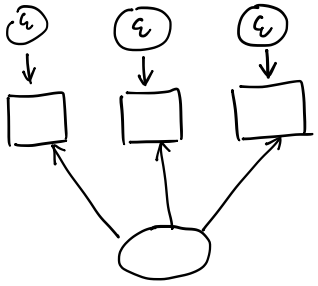
Structural equation modeling
- measurement model
- structural.

Variables
- observed variables
- latent variables
- endogenous variables
- exogenous variables.

SEM examines plausability of the hypothesized model.

If a model fits well, it just means that it's plausible.

But there may well be equivalent or better models



latent variables are unobserved.
∴ we need to fix its scale and location. For example,
1) fix one loading to be 1 and intercept to be 0
2) fix latent variable's variance directly and specify its mean.

This model would imply a set of variance and covariance for the observed variables. We can compare the implied variance covariance matrix and the sample variance-covariance matrix. If close, the model fits well.

In fact, the maximum likelihood seeks to minimize the difference between the actual and the implied covariance matrix.

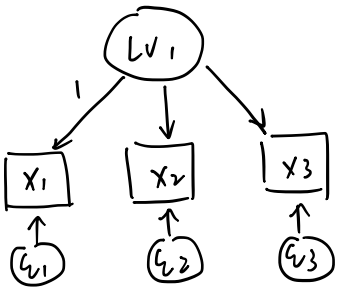
(if the model is saturated, there is no difference between the implied and the sample covariance matrix
→ there is only one saturated model and there is a close-form solution to the saturated model)

likelihood ratio test can be used to compare nested models

RMSEA ≤0.05 ✓ ⇒ closeness between the implied and the sample covariance.

CFI: compare the fitted model to a null model that assume no relationship.

Example 1



$$X_1 = v_1 + LV_1 + \varepsilon_1$$
$$X_2 = v_2 + \lambda_2 LV_1 + \varepsilon_2$$
$$X_3 = v_3 + \lambda_3 LV_1 + \varepsilon_3$$

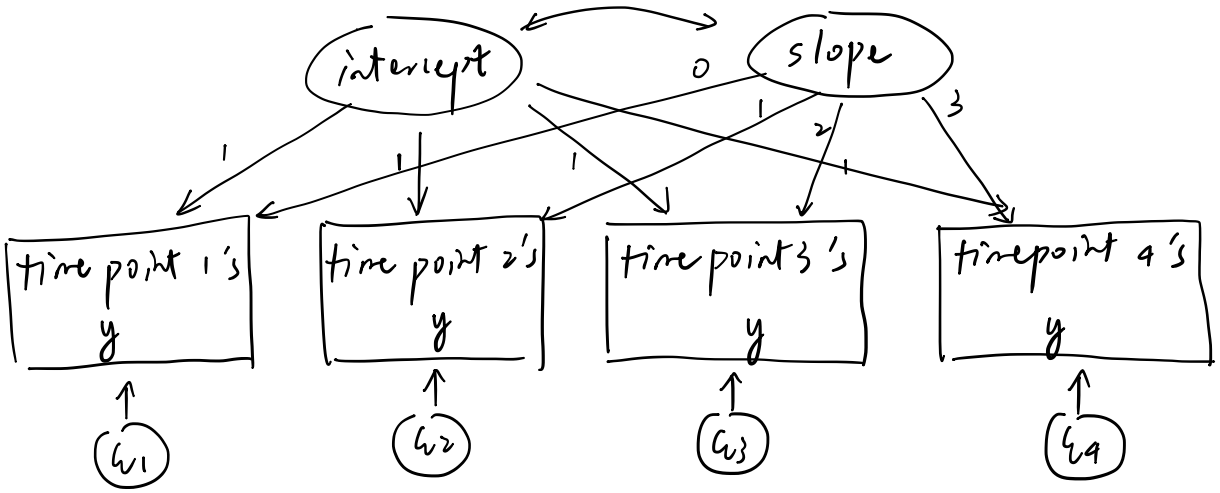
$$LV_1 \sim N(0, \sigma^2)$$
$$\begin{cases} \varepsilon_1 \sim N(0, \sigma_1^2) \\ \varepsilon_2 \sim N(0, \sigma_2^2) \\ \varepsilon_3 \sim N(0, \sigma_3^2) \end{cases}$$

$$\hat{E}(x) = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}$$

$$\hat{V}(x) = \begin{bmatrix} V(LV_1) + V(\varepsilon_1) & & \\ \lambda_2 V(LV_1) & \lambda_2^2 V(LV_1) + V(\varepsilon_2) & \\ \lambda_3 V(LV_1) & \lambda_2 \lambda_3 V(LV_1) & \lambda_3^2 V(LV_1) + V(\varepsilon_3) \end{bmatrix}$$

Example 3 latent growth curve model.

wide data. the information on time points are incorporated through the loadings that we specify



For individual i

time point 1's y_i = $\text{intercept}_i + 0 \cdot \text{slope}_i + \varepsilon_1$

time point 2's y_i = $\text{intercept}_i + 1 \cdot \text{slope}_i + \varepsilon_2$

↑ fix intercepts to 0

allows different people to have different starting points and different growth rates.

multi group analysis.

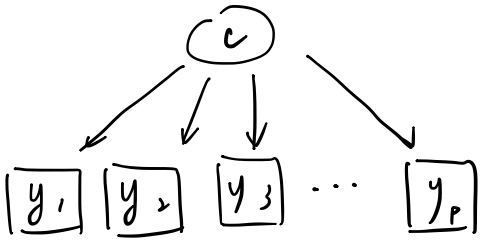
A different way to test / specify interactions.

measurement invariance

starting from free model, constrain the parameters to be the same across groups.

If the constraints do not significantly reduce model fits the invariance can be established.

latent class analysis.



joint probability of p items [← this is the model we specify]

$$P(y_1 y_2 \cdots y_p) = \sum^{K} P(c=k)\, P(y_1 y_2 \cdots y_p \mid c=k)$$ Assuming conditional independence.

$$= \sum^{k} P(c=k)\, P(y_1 \mid c=k)\, P(y_2 \mid c=k) \cdots P(y_p \mid c=k)$$

posterior probability

$$P(c=k \mid y_1 y_2 \cdots y_p) = \frac{P(y_1 y_2 \cdots y_p \mid c=k)\, P(c=k)}{P(y_1 y_2 \cdots y_p)}$$

How many classes in C is what we need to figure out.

likelihood ratio test can compare model with k and k-1 classes.

(doesn't have a χ^2 distribution.

use bootstrap to get an empirical distribution of LRT statistics)

At least 3 binary indicators are needed for a 2-class model.

with 2 binary indicators: 4 probabilities $P(x_1=0\ x_2=0)$, $P(x_1=0\ x_2=1)$, $P(x_1=1\ x_2=0)$, $P(x_1=1\ x_2=1)$

what we are estimating:

$P(x_1=1 \mid c=1)$
$P(x_2=1 \mid c=1)$
$P(x_1=1 \mid c=0)$
$P(x_2=1 \mid c=0)$
$P(c=1)$

5 parameters. — not identified.

latent class analysis + distal outcomes

different ways to account for the uncertainty of class assignment.

- Muplus/Muthen has a multi-step approach that takes this uncertainty into account.

- Can make multiple assignments based on the estimated probabilities and use the multiple imputation approach to incorporate the uncertainty.