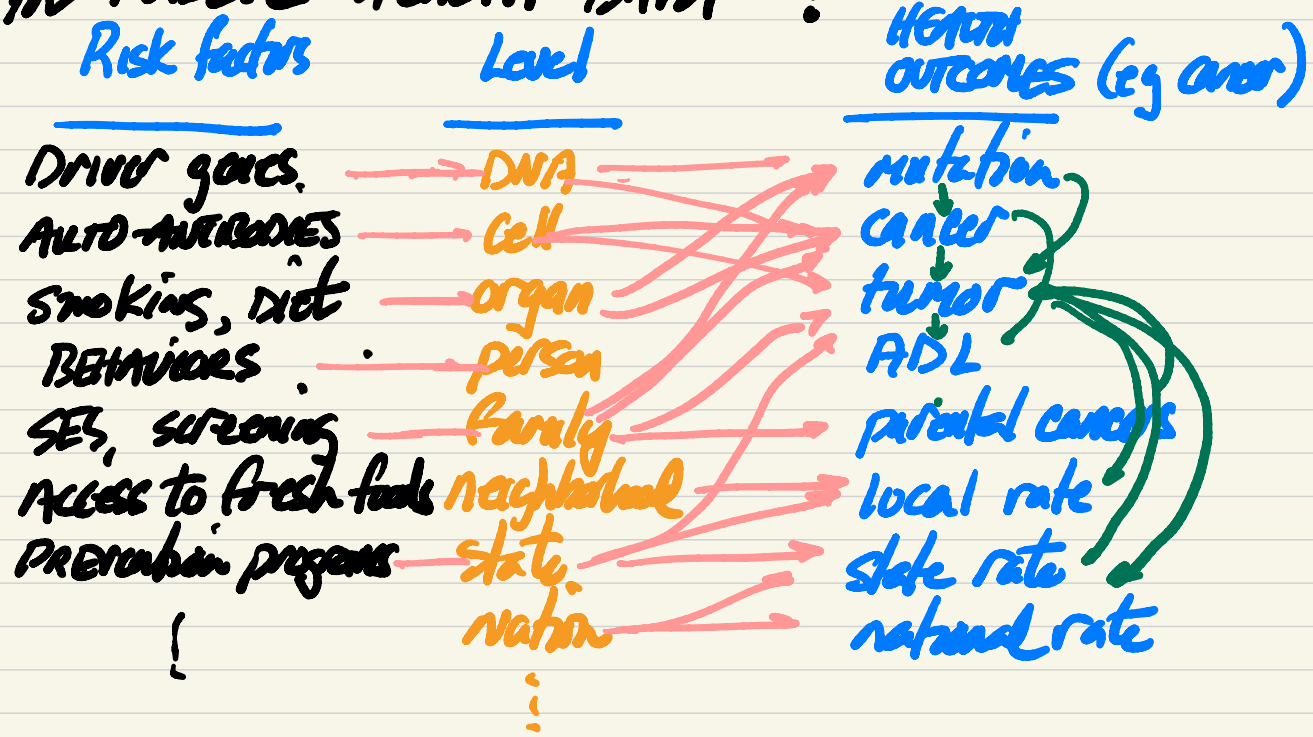
Multi-Level Models (MLMS) 2021

Final Questions for Discussion

1. WHAT IS THE MULTI-LEVEL (CAUSAL) MODEL in PUBLIC HEALTH DATA ?



Key ideas

- risk factors and outcomes occur at different spatial + temporal scales that interact with one another
- interventions targetted to levels : eg smoking : smoker education or patch -vs- no smoking in public/private buildings
- decompose evidence about relationship X → Y into different level-specific components; combine when appropriate

- clustered responses $(Y_{ij}, Y_{ij'})$ are likely correlated; correlation must be modelled to obtain:
  - valid inferences
  - efficient estimators
  - robustness to missing data patterns
- there are heterogeneities in covariate effects among clusters at each level of interest that give rise to the within-cluster correlations

2. WRITE and INTERPRET an INTERESTING ("RICH") 2-LEVEL MODEL TO ADDRESS THE QUESTION: does an 11-year olds' family SES affect his or her performance on a reading comprehension test and is the SES effect the same for boys and girls?

Y_{ij} - reading test score for child $j = 1, \ldots, n_i$ in school $i = 1, \ldots, m$

F_{ij} = 1 if female; 0 if male

S_{ij} = SES level

$$Y_{ij} = \beta_{0i} + \beta_{1i}(S_{ij} - \bar{S}_{..}) + \beta_2 F_{ij} + \beta_3 (S_{ij} - \bar{S}_{..}) F_{ij} + \varepsilon_{ij}$$

$$\begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} \sim G\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, D_{2\times 2} \right), \quad \varepsilon_{ij} \overset{iid}{\sim} G(0, \sigma^2)$$

$\begin{pmatrix}\beta_{0i}\\ \beta_{1i}\end{pmatrix}$ - $\begin{pmatrix}\text{intercept}\\ \text{S slope}\end{pmatrix}$ that define the linear dependence of reading on SES within all boys school i

β_2 - average difference in reading score comparing girls to boys of the same SES from same school

β_3 - the difference in the average SES effect on reading between girls and boys from same school

$D = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}$: $d_{11} = \text{Var}(\beta_{0i}) \Rightarrow \pm 2\sqrt{d_{11}}$ will include ~95% of intercepts; $d_{22} = \text{Var}(\beta_{1i}) \Rightarrow$ $\pm 2\sqrt{d_{22}}$ will include 95% of Y-S slopes around their mean slope that depends on $\bar{F}_{i\cdot}$

σ^2 - residual variance ($\text{Var}\,\varepsilon_{ij}$)

3. Continuing from (2), does the contextual effect of the average School SES differ for boys and girls?

$$Y_{ij} = \beta_{0i} + \beta_1 (S_{ij} - \bar{S}_{..}) + \beta_{1C} (\bar{S}_{i.} - \bar{S}_{..}) + \beta_2 F_{ij} + \beta_3 (\bar{S}_{i.} - \bar{S}_{..}) F_{ij} + \varepsilon_{ij}$$

φ	Contextual effect
Yes	$\beta_{1C} + \beta_3$
NO	β_{1C}

4. Consider a 3-level design in which children (k) within families (j) within villages (i) have lower respiratory infection ($Y_{ijk}=1$) or not ($Y_{ijk}=0$). GUESSTIMATE and Interpret the log odds ratio $(Y_{ijk}, Y_{i'j'k'})$ when:

		$\log OR(Y_{ijk}, Y_{i'j'k'})$
4.1	$i=i'$, $j=j'$, $k \neq k'$	$\alpha_0 + \alpha_1$
4.2	$i=i'$, $j \neq j'$, $k \neq k'$	α_0
4.3	$i \neq i'$ $j \neq j'$, $k \neq k'$	0

$$\log OR(Y_{ijk}, Y_{i'j'k'}) \quad \alpha_0 1_{\{i=i'\}} + \alpha_1 1_{\{i=i', j=j'\}}$$

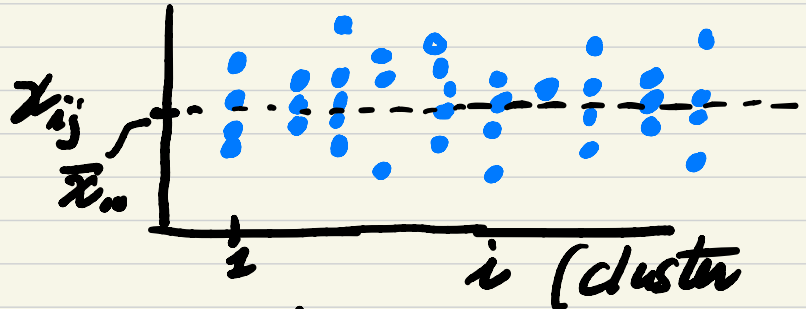
5. Consider the 2-level model:

$$Y_{ij} = \beta_{0i} + \beta_{1W}(x_{ij} - \bar{x}_{i\cdot}) + \beta_{2B}\bar{x}_{i\cdot} + \varepsilon_{ij}$$

$$\beta_{0i} = \beta_0 + b_i \; ; \quad b_i \overset{iid}{\sim} G(0, \tau^2), \; \varepsilon_{ij} \overset{iid}{\sim} G(0, \sigma^2)$$

$$Cov(b_i, \varepsilon_{ij}) = 0, \text{ for all } i, j$$

and the data set shown below



$$Var(\hat{\beta}_{1W}) \underset{(a)}{<=} Var(\hat{\beta}_{1B}) \quad ; \quad \underset{(a)}{\leq}, \underset{(b)}{=}, \underset{(c)}{\geq}, \underset{(d)}{NA}$$

EXPLAIN

6. If $\hat{\beta}_{1W} \neq \hat{\beta}_{1b}$ above, why?

- $\beta_{1W} = \beta_{1b}$ and unmeasured confounders
  - within cluster $\Rightarrow$ $\hat{\beta}_{1b}$ unbiased; $\hat{\beta}_{1W}$ biased
  - among clusters $\Rightarrow$ " biased; " unbiased
- $\beta_{1W} \neq \beta_{1b}$

7. What is the difference between a "marginal" or "population-average" MLM and a "conditional" or "cluster-specific" MLM? When does the distinction matter?

Marginal

$\log \text{odds}(Y_{ij}=1) = x_{ij}\beta^M$

Conditional given $\underset{\sim}{b}_i$:

$\log \text{odds}(Y_{ij}=1 \mid \underset{\sim}{b}_i) = x_{ij}\beta + z_{ij}\underset{\sim}{b}_i$

$|\beta^M| \leq |\beta^C|$

attenuation

8. In questions (2) and (3) above, how would you find out whether the sex-specific contextual effects estimates might be biased by the school compositions with respect to variables $\underset{\sim}{z}_{ij}$?

$$Y_{ij} = \beta_{0i} + \beta_{1i}(S_{ij} - \bar{S}_{\cdot\cdot}) + \beta_{1c}(\bar{S}_{i\cdot} - \bar{S}_{\cdot\cdot}) + \beta_2 F_{ij} + \beta_3 (\bar{S}_{i\cdot} - \bar{S}_{\cdot\cdot}) F_{ij} + \varepsilon_{ij}$$

$$\Downarrow$$

$$\beta_{0i} = \beta_0 + \beta_4 z_{ij}$$

How do β_{1c}, $\beta_{1c} + \beta_3$, $\mathrm{Var}\begin{pmatrix}\beta_{0i} \\ \beta_{1i}\end{pmatrix}$ change ?

9. As an MPH culminating project, a student conducts a meta-analysis of the following data

STUDY (i)	Effect Estimate ($\hat{\theta}_i$)	$\hat{se}(\hat{\theta}_i)$
1	-2	2
2	0	1
3	2	$\frac{1}{2}$

Reproduce her forest plot below

FOREST PLOT

STUDY	$\hat{\theta}$	$\widehat{se}_{\hat{\theta}}$
A	-2	2
B	0	1
C	2	$1/2$

Pooled

Model: $\hat{\theta}_i = \theta_i + \varepsilon_i \quad , \quad \theta_i \overset{iid}{\sim} G(a, \tau^2), \; \varepsilon_i \sim G(0, \widehat{se}^2_{\hat{\theta}_i})$

Fixed: $\tau^2 = 0 \Longrightarrow \tilde{\theta}_{(0)} = [(1/2)^2 + (1)^2 + (1/2)^2]^{-1}\left[\left(\tfrac{1}{2}\right)^2(-2) + 1\cdot 0 + (1/2)^2(2)\right] =$

$= (21/4)^{-1}(1/2) = \frac{4}{21} \cdot \frac{1}{2} = 30/21 = 1\,9/7 = 1.43$

$\widehat{se}_{\hat{\theta}_{(0)}} = \left(1/[1/2^2 + 1^2 + (1/2)^2]\right)^{1/2} = .43$

RANDOM:

$$\tau^2 = 1: \quad \tilde{\theta}_{(1)} = \left[\frac{1}{5} + \frac{1}{2} + \frac{1}{54}\right]^{-1}\left[-\frac{2}{5} + \frac{2}{54}\right]$$

$$= 0.8$$

$$se_{\hat{\theta}_{(1)}} = \left(1/\left[\tfrac{1}{5} + \tfrac{1}{2} + \tfrac{1}{54}\right]\right)^{1/2} = .82$$

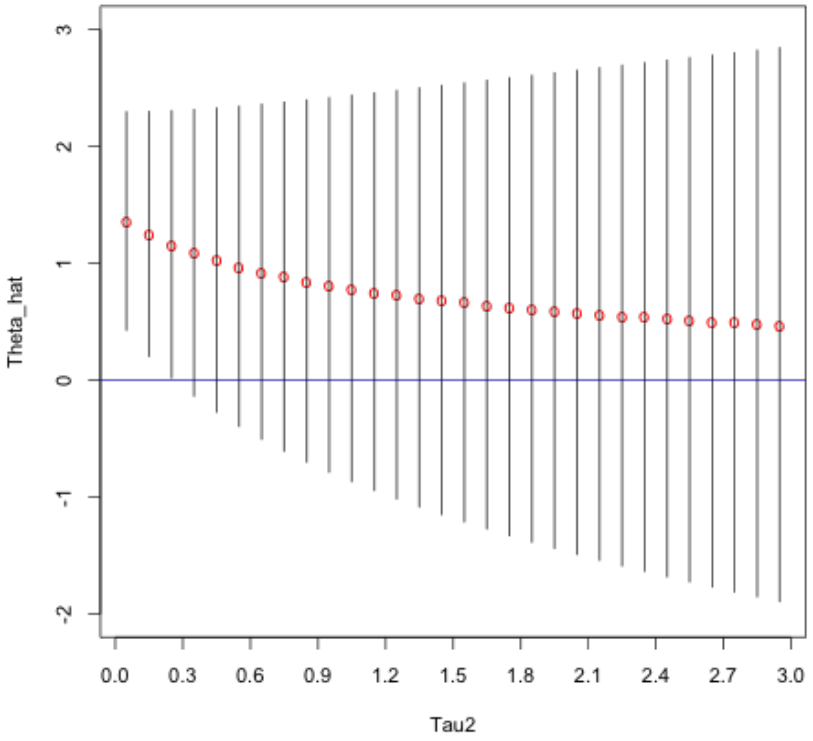
Can't trust estimate of τ^2. Use informative prior for τ^2, be Bayesian!

First, look at $\hat{\theta}_{(\tau^2)}$ for a range of plausible values of τ^2

Estimate $\hat{\tau}^2 = \frac{1}{2}\sum_{j=1}^{3}(\hat{\theta}_i - \bar{\hat{\theta}}_\cdot)^2 - \frac{1}{3}\sum_{i=1}^{3}(\widehat{se}_{\hat{\theta}_i})^2$

$= 2.25$

95% of true θ_i values range from $\bar{\theta}_\cdot \pm 4.5$? - scientifically plausible?

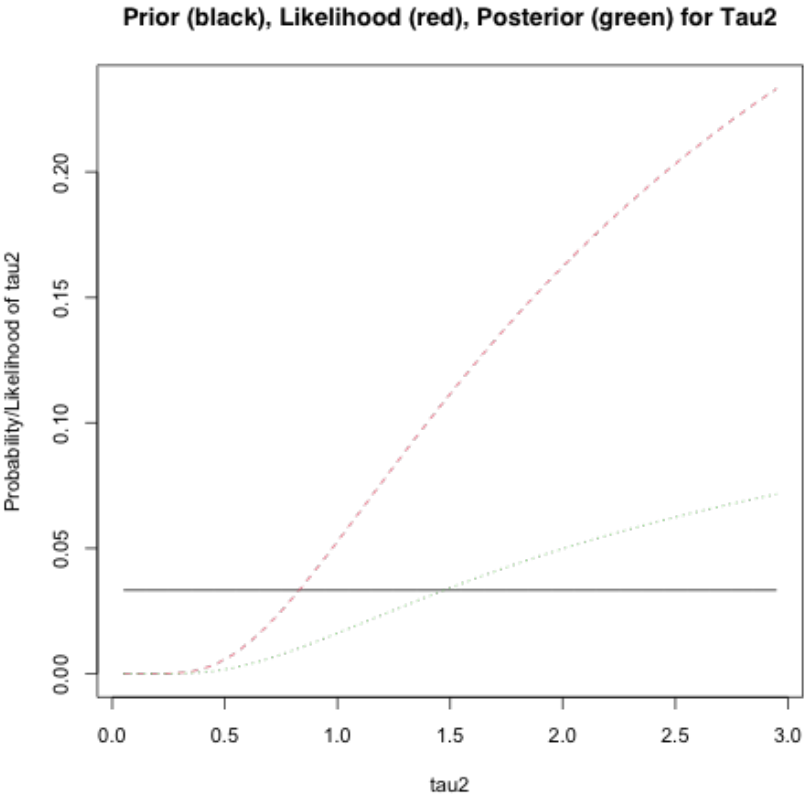

Theta_hat
3
2
1
0
-1
-2
0.0
0.3
0.6
0.9
1.2
1.5
1.8
2.1
2.4
2.7
3.0
Tau2

Prior for τ^2: $U(0,1)$

Likelihood of $\hat{\tau}^2$ given τ^2

$2\hat{\tau}^2/\tau^2 \sim \chi^2_2$

(in general $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$)



$\tilde{\tau}^2 \approx \sum \text{post}(\tau^2) \cdot \tau^2$

$= 2.13$

$\hat{\theta}(2.13) = 0.58_{-1.51}^{2.68}$

10. An HPM student is designing a 2-level study to estimate how the fraction of medical expenditures that are out-of-pocket (OOP) in a year differ between those families in insurance plans A and B. The student has #105,000. It costs #5,000 per cluster and #500 per person in the clusters to collect the data. She is considering 4 designs:

	n_C villages	n_A subjects/village	n_B	n_T total	$n_C \times n_T$
I.	10	5	5	10	100
II.	10	0 (10)	10 (0)	10	100
III.	15	2	2	4	60
IV.	15	0 (4)	4 (0)	4	60

How would you advise her as the HPM expert?

Thoughts

- Compare A vs B within (I/III) or among clusters (II/IV)?

  depends on relative sizes of $Var(b_{0i})$, $Var(\epsilon_{ij})$

- Model (cartoon)

$$Plan_j = \begin{cases} 1 & A \\ 0 & B \end{cases}$$

$$Y_{ij} = \beta_{0i} + \beta_1 Plan_j + \epsilon_{ij}$$

$$\beta_{0i} \overset{iid}{\sim} G(\beta_0, \tau^2), \quad \epsilon_{ij} \overset{iid}{\sim} G(0, \sigma^2)$$

$$b_{0i} \perp \epsilon_{ij}$$

Within cluster comparison

$$Var(\bar{y}_{iA} - \bar{y}_{iB}) = Var(\beta_0 + \beta_1 + \bar{\varepsilon}_{iA} - \beta_0 - \bar{\varepsilon}_{iB})$$

$$= \boxed{\frac{\sigma^2}{n_a} + \frac{\sigma^2}{n_b}}$$

Across cluster comparison i - gets A, i' gets B

$$Var(\bar{y}_{iA} - \bar{y}_{i'B}) \quad i \neq i' \quad =$$

$$Var(= \beta_0 + \beta_1 + \bar{\varepsilon}_{iA} - \beta_0 + \bar{\varepsilon}_{iB})$$

$$\boxed{= 2\tau^2 + \frac{\sigma^2}{n_a} + \frac{\sigma^2}{n_b}}$$

Designs

I/III : $\text{Var}\,\hat{\beta}_1 = \frac{1}{m}\text{Var}(\bar{y}_{A\cdot} - \bar{y}_{B\cdot}) = \frac{\sigma^2}{m}\left(\frac{1}{n_a} + \frac{1}{n_b}\right)$

II/IV : $\text{Var}\,\hat{\beta}_1 = \frac{2\tau^2}{m/2} + \frac{\sigma^2}{\frac{m}{2}}\left(\frac{1}{2n_a} + \frac{1}{2n_b}\right)$

Note in designs II/IV, we have half as many clusters $\left(\frac{m}{2}\right)$ getting each plan but twice as many persons/cluster $(2n)$

Also assume $n_a = n_b = n$ for simplicity

$$\text{I/III}: \quad \text{Var}\,\hat{\beta}_1 = \frac{\sigma^2}{m}\left(\frac{2}{n}\right) = \frac{2\sigma^2}{mn}$$

$$\text{II/IV}: \quad \text{Var}\,\hat{\beta}_1 = \frac{4\tau^2}{m} + \frac{2\sigma^2}{mn}$$

$$\text{Efficiency} = \frac{\text{Var}\,\hat{\beta}_1(\text{I/III})}{\text{Var}\,\hat{\beta}_1(\text{II/IV})} = \frac{2\sigma^2/(mn)}{\frac{4\tau^2}{m} + \frac{2\sigma^2}{mn}}$$

$$= \frac{1}{2\frac{4\tau^2}{m}\cdot\frac{mn}{2\sigma^2} + 1} = \boxed{\frac{1}{2\frac{n\tau^2}{\sigma^2} + 1}}$$

$\tau^2 = 0 \Rightarrow$ no correlation within clusters, no benefit to within-cluster comparison.

$\tau^2 > 0$, design I/III favored more as $n \uparrow$

	m	n_a+n_b	τ^2	σ^2	Efficiency
A-vs-B	10	10	0	1	1.0
			1	4	$0.1\overline{6}$
			4	1	0.03
C-vs-D	15	4	0	1	1.0
			1	4	$0.3\overline{3}$
			4	1	0.08