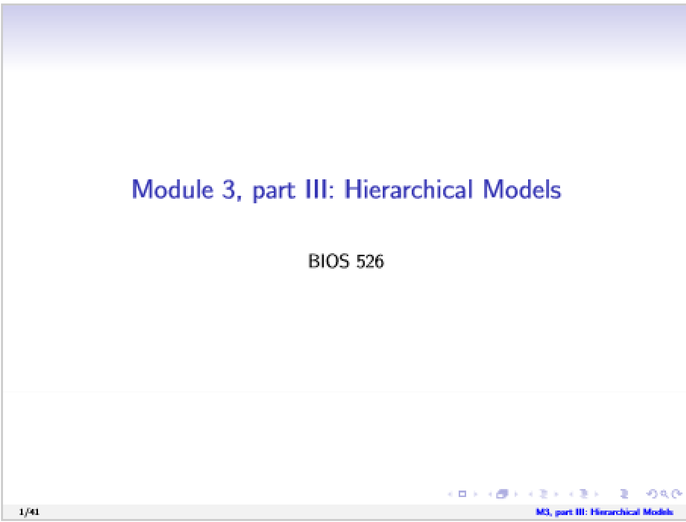
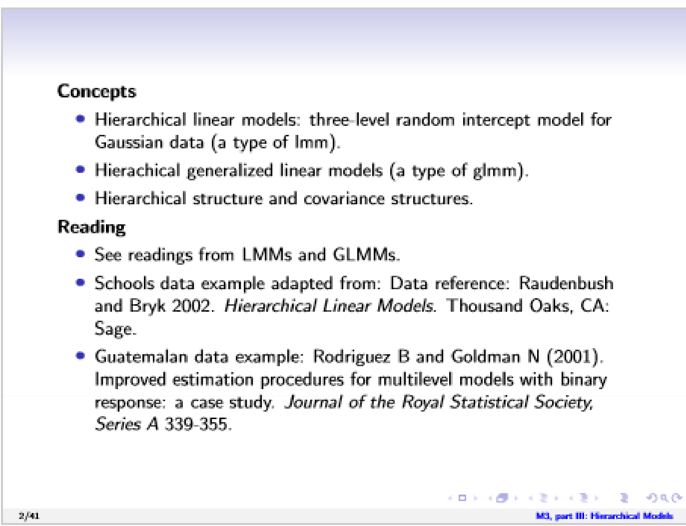
Module 3 Part 3: Hierarchical Models

Monday, September 18, 2023 14:24

BIOS526_M
3_PartIII_...





Hierarchical Linear Models

Math Achievement Data

- Longitudinal study of children's academic growth.
- 1,721 students from 60 urban primary schools.
- Standardized math achievement scores recorded at each primary school year (1-6).
- Scientific question: what child-level and school-level factors influence academic growth?
- Outcome data y_{ijk} has three levels:

Level 3: School $i = 1, \ldots, 60$
Level 2: Child $j = 1, \ldots, 1721$
Level 1: Yearly math scores $k = 1, \ldots, 6$

The above multi-level data have a nested structure because the clusters (*child*) are themselves grouped within *school*.

Math Achievement Data: Variables

i^{th} school
j^{th} child
k^{th} year

Level 1 (year within child): finest level: values vary with k

- $math_{ijk}$: math scores (outcome)
- $year_{ijk}$: primary school year centered at 3.5
- $retained_{ijk}$: indicator for child ij repeating the grade in year k.

Level 2 (child within school): values vary with j (constant over k)

- $child_{ij}$: child ID
- $female_{ij}$: indicator for child i in school j being female
- $black_{ij}$: indicator for child i in school j being African American
- $hispanic_{ij}$: indicator for child i in school j being Hispanic

Level 3 (school): values vary with i: highest (most granular) level (constant over j and k)

- $school_i$: school ID
- $size_i$: number of students in school i
- $lowinc_i$: percent of students from low-income families in school i

Three-Level Normal Random Intercept Model

Level 1: relating math scores to occasion (year) specific covariates.

↳ fast varying index

$$math_{ijk} = \beta_{0,ij} + \beta_1 year_{ijk} + \beta_2 retained_{ijk} + \epsilon_{ijk}$$

collects everything at level 2

↳ everything else thrown into here

Level 2: relating child-specific random intercepts to child characteristics.

$$\beta_{0,ij} = \alpha_{0,i} + \alpha_1 female_{ij} + \alpha_2 black_{ij} + \alpha_3 hispanic_{ij} + \psi_{ij}$$

collects everything at level 3

these are all constant between years

↳ child random effect

Level 3: relating school-specific random intercepts to school characteristics.

$$\alpha_{0,i} = \gamma_0 + \gamma_1 size_i + \gamma_2 lowinc_i + \eta_i$$

→ school-random effect

$$\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2) \qquad \psi_{ij} \overset{iid}{\sim} N(0, \tau^2) \qquad \eta_i \overset{iid}{\sim} N(0, \nu^2)$$

$$\epsilon_{ijk} \perp\!\!\!\perp \psi_{ij} \perp\!\!\!\perp \eta_i$$

all these random effects are independent

Three-Level Normal Random Intercept Model

Level 1: $math_{ijk} = \beta_{0,ij} + \beta_1 year_{ijk} + \beta_2 retained_{ijk} + \epsilon_{ijk}$

- A child's score in year k is a linear function of $year_{ijk}$ and $retained_{ijk}$.
- $\beta_{0,ij}$ is the child-specific random intercept. Note that the coefficient has subscript ij as it cannot be influenced by variables that change between school years.

  child specific effect
- Between-child variation is accounted for by $\beta_{0,ij}$.
- After removing the child-specific intercept, residual variation in math scores follow $N(0, \sigma^2)$.

  and year & retained effects

Three-Level Normal Random Intercept Model

Level 2: $\beta_{0,ij} = \alpha_{0,i} + \alpha_1 female_{ij} + \alpha_2 black_{ij} + \alpha_3 hispanic_{ij} + \psi_{ij}$

- This second-level regression model explains variation in $\beta_{0,ij}$.
- We assume $\beta_{0,ij}$ is a linear function of *child-specific covariates* $female_{ij}$, $black_{ij}$, and $hispanic_{ij}$.
- $\alpha_{0,i}$ is a school-level random intercept. It captures correlation in $\beta_{0,ij}$ for children from the same school.
- Variation in child-specific $\beta_{0,ij}$ not explained by child-level covariates and the school-level intercepts follow $N(0, \tau^2)$.

Three-Level Normal Random Intercept Model

Level 3: $\alpha_{0,i} = \gamma_0 + \gamma_1 size_i + \gamma_2 lowinc_i + \eta_i$

- Level 3 explains variation in school-specific intercepts $\alpha_{0,i}$ using school-level covariates $size_i$ and $lowinc_i$.
- We assume $\alpha_{0,i}$ is normal with variance ν^2.
- γ_0 is the *overall* baseline mean math score across 60 schools. Here baseline is 0% low-income students and zero students.
- γ_1 can be interpreted as: increase in school-average math score per unit increase in $size_i$.

$\eta_i \overset{\text{IID}}{\sim} N(0, \nu^2)$

Three-Level Normal Random Intercept Model

The multilevel model can be combined to give:

$$
\begin{aligned}
math_{ijk} = {} & \gamma_0 + \gamma_1 size_i + \gamma_2 lowinc_i + \eta_i \\
& + \alpha_1 female_{ij} + \alpha_2 black_{ij} + \alpha_3 hispanic_{ij} + \psi_{ij} \\
& + \beta_1 year_{ijk} + \beta_2 retained_{ijk} + \epsilon_{ijk}
\end{aligned}
$$

$$\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2) \qquad \psi_{ij} \overset{iid}{\sim} N(0, \tau^2) \qquad \eta_i \overset{iid}{\sim} N(0, \nu^2)$$

- Because $size_i$ and $lowinc_i$ are the same values for all scores taken in a particular school, γ_1 and γ_2 change $math_{ijk}$ on the school-level. Every measurement and every child in school i has the same $\gamma_1 size_i + \gamma_2 lowinc_i + \eta_i$.
- γ_0 is the overall mean math score at baseline:
  - $size_i = 0$, $lowinc_i = 0$
  - male, non-black, non-hispanic
  - grade year at 3.5, not retained

Variances

$$
\begin{aligned}
math_{ijk} = {} & \gamma_0 + \gamma_1 size_i + \gamma_2 lowinc_i + \eta_i \\
& + \alpha_1 female_{ij} + \alpha_2 black_{ij} + \alpha_3 hispanic_{ij} + \psi_{ij} \\
& + \beta_1 year_{ijk} + \beta_2 retained_{ijk} + \epsilon_{ijk}
\end{aligned}
$$

$$\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2) \qquad \psi_{ij} \overset{iid}{\sim} N(0, \tau^2) \qquad \eta_i \overset{iid}{\sim} N(0, \nu^2)$$

Then $= Var(\epsilon_{ijk} + \psi_{ij} + \eta_i)$ and since $\perp\!\!\!\perp$ ...

$$Var(math_{ijk}) = Var(\eta_i) + Var(\psi_{ij}) + Var(\epsilon_{ijk}) = \nu^2 + \tau^2 + \sigma^2$$

Covariances and Intra-Subject Correlation

same i, same j, diff k (k vs. k')

$$
\begin{aligned}
Cov(math_{ijk}, math_{ijk'}) &= Cov(\eta_i + \psi_{ij} + \epsilon_{ijk}, \eta_i + \psi_{ij} + \epsilon_{ijk'}) \\
&= Cov(\eta_i + \psi_{ij}, \eta_i + \psi_{ij}) \\
&= \tau^2 + \nu^2
\end{aligned}
$$

Within-child correlation = correlation between different scores within the same child (must be within the same school):

$$
Cor(math_{ijk}, math_{ijk'}) = \frac{\tau^2 + \nu^2}{\tau^2 + \nu^2 + \sigma^2}
$$

Intra-School Correlation

same i, diff j, diff k

Then for different students at same school,

$$
\begin{aligned}
Cov(math_{ijk}, math_{ij'k'}) &= Cov(\eta_i + \psi_{ij} + \epsilon_{ijk}, \eta_i + \psi_{ij'} + \epsilon_{ij'k'}) \\
&= Cov(\eta_i, \eta_i) \\
&= \nu^2
\end{aligned}
$$

Within-school correlation = correlation between different scores within the same school (different child):

$$
Cor(math_{ijk}, math_{ij'k'}) = \frac{\nu^2}{\sigma^2 + \tau^2 + \nu^2}
$$

Note that within-child scores are more similar than within-school scores.

Note that in a nested-structure, we don't have data from the same child but different schools.

G is an example of y_{ijk} and $y_{i'jk}$ are different students

Covariance Matrix

$Cov(y_{ijk}, y_{ij'k}) = Cov(y_{ijk}, y_{ij'k'})$ here bc years no diff??

School i	Student j	Year k	111	112	113	121	122	123	211	212	213
1	1	1	$\boxed{\sigma^2+\tau^2+\nu^2}$	$\tau^2+\nu^2$	$\tau^2+\nu^2$	γ^2	γ^2	γ^2	0	0	0
1	1	2	$\tau^2+\nu^2$	$\boxed{\sigma^2+\tau^2+\nu^2}$	$\tau^2+\nu^2$	γ^2	γ^2	γ^2	0	0	0
1	1	3	$\tau^2+\nu^2$	$\tau^2+\nu^2$	$\boxed{\sigma^2+\tau^2+\nu^2}$	γ^2	γ^2	γ^2	0	0	0
1	2	1	ν^2	ν^2	ν^2				0	0	0
1	2	2	γ^2	γ^2	ν^2				0	0	0
1	2	3	γ^2	γ^2	γ^2				0	0	0
2	1	1	0	0	0	0	0	0			
2	1	2	0	0	0	0	0	0			
2	1	3	0	0	0	0	0	0			

Hierarchical Specification

$$math_{ijk} = \beta_{0,ij} + \beta_1 year_{ijk} + \beta_2 retained_{ijk} + \epsilon_{ijk}$$
$$\beta_{0,ij} = \alpha_{0,i} + \alpha_1 female_{ij} + \alpha_2 black_{ij} + \alpha_3 hispanic_{ij} + \psi_{ij}$$
$$\alpha_{0,i} = \gamma_0 + \gamma_1 size_i + \gamma_2 lowinc_i + \eta_i$$
$$\epsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2) \qquad \psi_{ij} \overset{iid}{\sim} N(0, \tau^2) \qquad \eta_i \overset{iid}{\sim} N(0, \nu^2)$$

The above model also says:

given ψ_{ij}, η_i

$$math_{ijk} \sim N\left(\beta_{0,ij} + \beta_1 year_{ijk} + \beta_2 retained_{ijk},\ \sigma^2\right)$$
$$\beta_{0,ij} \sim N\left(\alpha_{0,i} + \alpha_1 female_{ij} + \alpha_2 black_{ij} + \alpha_3 hispanic_{ij},\ \tau^2\right)$$
$$\alpha_{0,i} \sim N\left(\gamma_0 + \gamma_1 size_i + \gamma_2 lowinc_i,\ \nu^2\right)$$

M3_partIII_HM.R

He shows how to check the data

Data Example

```
> dat = read.csv ("achievement.csv")

> dim (dat)
[1] 7230   10

> dat[1:10,]
   year   math retained female black hispanic size lowinc school child
1   0.5  1.146        0      0     0        1  380   40.3   2020   244
2   1.5  1.134        0      0     0        1  380   40.3   2020   244
3   2.5  2.300        0      0     0        1  380   40.3   2020   244
4  -1.5 -1.303        0      0     0        0  380   40.3   2020   248
5  -0.5  0.439        0      0     0        0  380   40.3   2020   248
6   0.5  2.430        0      0     0        0  380   40.3   2020   248
7   1.5  2.254        0      0     0        0  380   40.3   2020   248
8   2.5  3.873        0      0     0        0  380   40.3   2020   248
9  -1.5 -1.384        0      0     0        1  380   40.3   2020   253
10 -0.5  0.338        0      0     0        1  380   40.3   2020   253

> length (unique (dat$child))
[1] 1721

> length (unique (dat$school))
[1] 60
```


stuff like this could be on exam

Load lmerTest

Model Fitting and Interpretations

We specify the multi-level model with two random intercept components. (Here, child ID is unique, so the below code is equivalent to (1|school)+(1|school:child), see R code.) → if don't have a unique id → check M3_partIII_... for more info

```
> fit = lmer (math ~ year + retained + female + black + hispanic + size + lowinc
                     + (1|school) + (1 | child), data = dat)

> summary (fit)

Random effects:
 Groups   Name        Variance Std.Dev.
 child    (Intercept) 0.663673 0.81466
 school   (Intercept) 0.087498 0.29580
 Residual             0.344524 0.58696
```

Grand intercept for school

Grand intercept for child

child ID unique here so it doesn't matter how you specified

$\hat{\tau}^2$ $\hat{\nu}^2$ $\hat{\sigma}^2$

Note that the random effects of *child* and *school* are assumed to be independent.

- Within-child correlation $= \frac{0.66+0.087}{0.66+0.087+0.34} = 0.69$
- Within-school correlation $= \frac{0.087}{0.66+0.087+0.34} = 0.08$

if your data is coded weird (like nested IDs) you can do school : child

reference Bates table?

Nesting versus crossed

If two factors are crossed, all levels of one factor appear in all levels of the other factor. You can tell if two factors are crossed using a cross tabulation:

```
> table(dat$female, dat$black)
     0    1
0 1119 2426
1 1134 2551
```

If two factors are nested, then levels of one factor only appear in one level reference Bates table?

Nesting versus crossed

If two factors are crossed, all levels of one factor appear in all levels of the other factor. You can tell if two factors are crossed using a cross tabulation:

```
> table(dat$female, dat$black)
     0    1
  0 1119 2426
  1 1134 2551
```

If two factors are nested, then levels of one factor only appear in one level of another factor:

```
> table(dat$child,dat$school)
```

	2020	2040	2180	2330	2340	2380	2390	2440	2480	2520	2540	2560	2610	2620	2750	2820
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0

Nesting versus crossed is determined by the study design.

Note on nesting → determined from experimental design

A child is nested within school because for a given child, school does not vary.

Therefore, we can't look at the interaction between child and school. One could hypothesize that the same child at different schools creates additional variability, but we can't examine that here.

We can only look at interactions for crossed factors.

We can have cross-level interactions. We need to be able to observe the different combinations of the effect levels.

An interaction between different levels results in a variable of the finer level. E.g., does sex modify the effect of retention?

There are many possible interactions to consider here, but to keep things simple, we will ignore them.

e.g. does retention differ between boys & girls?

Model Fitting and Interpretation

fixed effects

3.5 baseline

```
Fixed effects:
                 Estimate Std. Error        df t value Pr(>|t|)
(Intercept)     2.398e-01  1.524e-01 6.406e+01   1.573  0.12064
year(grade) ijk 7.483e-01  5.396e-03 5.744e+03 138.685  < 2e-16 ***
retained ijk    1.481e-01  3.535e-02 5.802e+03   4.190 2.83e-05 ***
female ij      -9.038e-05  4.223e-02 1.668e+03  -0.002  0.99829
black ij       -5.182e-01  8.060e-02 1.154e+03  -6.429 1.88e-10 ***
hispanic ij    -2.899e-01  8.910e-02 1.642e+03  -3.254  0.00116 **
size i         -1.028e-04  1.485e-04 5.719e+01  -0.692  0.49167
lowinc i       -8.002e-03  1.818e-03 6.900e+01  -4.401 3.84e-05 ***
```

- Across schools, children, and school years, the average math score is 0.2398 for baseline measurement (year = 3.5, retained=0, female=0, black=0, hispanic=0, size=0, lowinc=0, ave school and ave child). Intercept is not meaningful since it is for size=0 and has large SE (see R code) → if interested in baseline, can be better to center
- see next slide [Lower child-specific average math scores were associated with African American and Hispanic students.
- Lower school-specific average math scores were associated with schools with higher proportion of low-income students. Low income significant
- Math scores increased as a child progressed in grade.
- Math scores higher for grades that a child repeated (retained).

Study Limitations

→ here, is school level, not individual level

- `lowinc` is the proportion of students at a school that are lower income (Level 3 covariate).
- We can not estimate whether a student from a disadvantaged background (low income, single-family household, parents' education, others) has lower scores.
- In particular, race and ethnicity are correlated with other factors impacting an individual's achievement scores.
- The lower scores reflect products of systemic racism and shortcomings of the education system. Institutional racism
- A brief description of systemic racism: https://www.youtube.com/watch?v=YrHIQIO_bdQ Structural racism
- Results of this data set could be used to help prioritize education funds or policy.

Comparison with 2-Level Models

Our 3-level model decomposes the total residual variation into three components:

- Level 3: Between school variation ν^2
- Level 2: Between child variation τ^2
- Level 1: Between grade variation (within a child) σ^2.

We can consider models that only include random intercepts for schools or for children.

Assume no between-child variation: only include $\eta_i \sim N(0, \nu^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$

```
> fit2 = lmer (math ~ year + retained + female + black + hispanic + size + lowinc
                     + (1|school) , data = dat)
```

Assume no between-school variation: only include $\varphi_{ij} \sim N(0, \tau^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$

```
> fit3 = lmer (math ~ year + retained + female + black + hispanic + size + lowinc
                     + (1|child) , data = dat)
```


Selecting random effects

I suggest using a model that seems reasonable.

LRTs can be inaccurate due to issues of testing a null hypothesis on the boundary of the parameter space (generally, p-values too big).

↳ helpful guidance, but in general use your judgement

```
> anova(fit2,fit)
refitting model(s) with ML (instead of REML)
Data: dat
Models:
fit2: math ~ year + retained + female + black + hispanic + size + lowinc + (1 | school)
fit: math ~ year + retained + female + black + hispanic + size + lowinc +
fit:     (1 | school) + (1 | child)
     npar   AIC   BIC   logLik deviance Chisq Df Pr(>Chisq)
fit2   10 20459 20528 -10219.4    20439
fit    11 16673 16748  -8325.4    16651  3788  1  < 2.2e-16 ***
---

> anova(fit3,fit)
refitting model(s) with ML (instead of REML)
Data: dat
Models:
fit3: math ~ year + retained + female + black + hispanic + size + lowinc + (1 | child)
fit: math ~ year + retained + female + black + hispanic + size + lowinc +
(1 | school) + (1 | child)
     npar   AIC   BIC  logLik deviance Chisq Df Pr(>Chisq)
fit3   10 16773 16842 -8376.4    16753
fit    11 16673 16748 -8325.4    16651   102  1  < 2.2e-16 ***
```

Comparison with 2-Level Models

Point Estimates (Standard Error)

	Random Intercept Grouping			
	Parameter	School	Child	Both
	Intercept	0.26 (0.14)	0.31 (0.08)	0.24 (0.15)
	year	0.74 (0.01)	0.75 (0.005)	0.75 (0.005)
	retained	-0.49 (0.05) → but here, turns negative but still significant	0.14 (0.03)	0.15 (0.03) → significant here
	female	-0.02 (0.02)	0.02 (0.04)	0.00 (0.04)
	black	-0.52 (0.05)	-0.43 (0.01)	-0.52 (0.01)
	hispanic	-0.29 (0.05)	-0.26 (0.01)	-0.29 (0.01)
	size	-0.0001 (0.0001)	-0.00001 (0.000007)	-0.0001 (0.0001)
	lowinc	-0.007 (0.001)	-0.009 (0.001)	-0.008 (0.001)
school	ν^2	0.10		0.09
child	τ^2		0.75	0.66
error	σ^2	0.97	0.34	0.34

Comparison with 2-Level Models

School-only → Child-only

- Large reduction in residual error variance σ^2.
- Most statistically significant coefficients have smaller standard errors.
- The effect of *retain* changes direction!
  - Children who were retained in a grade were associated with lower scores in school-only model. This child-specific effect is not controlled for by school-level intercepts.

Child-only → Both

- Part of the between-child variation is allocated to between-school variation: $0.75 = 0.09 + 0.66$.
- Residual errors and coefficient standard errors nearly unchanged.
- Overall intercept decreases from 0.31 to 0.24. For child-only, the intercept is the baseline (see previous) average score for a *typical child*. For the full model, intercept is baseline average score for *typical child* in *typical school*.

Nested versus crossed random effects

It is also possible for the random effects to be crossed. If the subject ID is the same at different schools, the following

```
(1|school)+(1|child)
```

will result in $Cov(y_{ijk}, y_{i'jk}) = \tau^2$.

It is possible for a subject ID to be coded poorly, such that ID 1 at school 1 corresponds to a different individual than ID 1 at school 2.

When this is the case, you must use the syntax

```
(1|school)+(1|school:child)
```

In our dataset, the correct covariance structure is used in lmer because the child ID is unique.

see R code

Hierarchical GLMMs

Guatemalan Vaccination Data

- Cross-sectional study of families' decision to immunize their children.
- Surveyed 1,595 mothers in 161 communities in Guatemala in 1987.
- Collected children immunizaton status born in the previous 5 years.
- Scientific question: whether a campaign in 1986 increased immunization rate. 1983 1984 1985 / 1986 1987
- Scientific question: were children at least 2-years old at time of interview more likely to be immunized?
- Outcome data y_{ijk} has three levels:

Level 3: Community i
Level 2: Family j
Level 1: Child k

The data are nested: family nested in community.

Guatemalan Vaccination Data: Variables

community mother child

Level 1 (kth Child within mother)

- $immun_{ijk}$: indicator for the child being immunized (outcome)
- $kid2p_{ijk}$: indicator for the child being at least 2-years-old at interview =1 if got vax info

Level 2 (jth family within community)

- mom_{ij}: mother's (family) ID
- $momEduPri_{ij}$: indicator for mother having primary education
- $momEduSec_{ij}$: indicator for mother having secondary education
- $husEduPri_{ij}$: indicator for husband having primary education
- $husEduSec_{ij}$: indicator for husband having secondary education

Level 3 (ith community)

- $cluster_i$: community ID → cluster = community
- $rural_i$: indicator for rural community
- $pcInd81_i$: percent population that was indigenous in 1981

Guatemalan Vaccination Data

```
> dat = read.csv ("guatemalan.csv")
> dim (dat)
[1] 2159   10

  mom cluster immun kid2p momEdPri momEdSec husEdPri husEdSec rural   pcInd81
1    2       1     1     1        0        1        0        1     0 0.1075042
2  185      36     0     1        1        0        1        0     0 0.0437295
3  186      36     0     1        1        0        0        1     0 0.0437295
4  187      36     0     1        1        0        1        0     0 0.0437295
5  188      36     0     1        1        0        0        0     0 0.0437295
6  188      36     1     1        1        0        0        0     0 0.0437295
7  189      36     1     1        0        1        1        0     0 0.0437295
8  190      36     1     0        1        0        1        0     0 0.0437295
9  190      36     1     1        1        0        1        0     0 0.0437295
10 191      36     1     1        1        0        0        1     0 0.0437295

> length (unique (dat$mom)) #Total number of mothers
[1] 1595

> length (unique (dat$cluster)) #Total number of communities
[1] 161

> table ( table (dat$mom)) #Number of children per mom
   1    2    3
1063  500   32
```


Mom's ID Nested Within Community (Cluster)

```
### Demonstrate with first two mother's IDs
> table (dat$mom, dat$cluster)

      1 36 38 45 46 47 49 50 51 55
  2   1  0  0  0  0  0  0  0  0  0
  185 0  1  0  0  0  0  0  0  0  0

> table (dat$mom, dat$cluster)[1:2, ] != 0

          1    36    38    45    46    47    49    50    51    55
mom 2     TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
mom 185 FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

> apply( table (dat$mom, dat$cluster)[1:2, ] != 0, 1, sum)
  2 185
  1   1

### Now apply to all mothers' IDs
> table( apply( table (dat$mom, dat$cluster) != 0, 1, sum) )
   1
1595                    =1 if each mom is in 1 cluster

## Grouping IDs okay. Each mother's ID appears only once in each community.
## This gives us the desired "nested" structure.
```


Three-Level Logistic Random Intercept Model

$$immun_{ijk} \sim \text{Binomial}\ (p_{ijk})$$

$$\text{logit}\ (p_{ijk}) = \beta_0 + u_i + u_{ij} + \beta_1' \boldsymbol{x}_{ijk} + \beta_2' \boldsymbol{x}_{ij} + \beta_3' \boldsymbol{x}_i,$$

$\mu_{ij} \perp\!\!\!\perp \mu_i$

$$u_{ij} \overset{iid}{\sim} N(0, \tau^2) \qquad u_i \overset{iid}{\sim} N(0, \nu^2)$$

community level

mother level

- $\boldsymbol{x}_{ijk} = [kid2p_{ijk}]$
- $\boldsymbol{x}_{ij} =$ $[momEduPri_{ij}, momEduSec_{ij}, husEduPri_{ij}, husEduSec_{ij}]$
- $\boldsymbol{x}_i = [rural_i, pcInd81_i]$ community variable
- $\tau^2 =$ between-family variation in baseline log odds
- $\nu^2 =$ between-community variation in baseline log odds
- Only two normal random effects. → NOTE: NO EPSILON!!
- $Var(immun_{ijk} | u_i, u_{ij}) = p_{ijk}(1 - p_{ijk})$.
- We can think of u_i and u_{ij} as parameters that control for unmeasured confounders at the community- and family-level. One statistician's mean structure is another's covariance structure.

Three-Level Logistic Random Intercept Model

Recall the model is conditioned on the random effects:

$P_{ijk} = E[y_{ijk} \mid \mu_i, \mu_{ij}]$

$$immun_{ijk} \sim \text{Binomial}\,(E[y_{ijk}|u_i, u_{ij}])$$

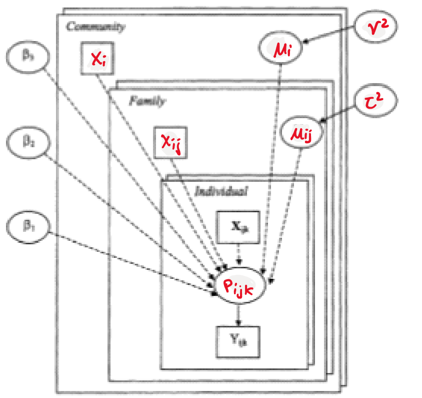
$$\text{logit}\,(E[y_{ijk}|u_i, u_{ij}]) = \beta_0 + u_i + u_{ij} + \beta_1' \boldsymbol{x}_{ijk} + \beta_2' \boldsymbol{x}_{ij} + \beta_3' \boldsymbol{x}_i,$$

$$u_{ij} \overset{iid}{\sim} N(0, \tau^2) \qquad u_i \overset{iid}{\sim} N(0, \nu^2)$$

- $\boldsymbol{x}_{ijk} = [kid2p_{ijk}]$
- $\boldsymbol{x}_{ij} = [momEduPri_{ij}, momEduSec_{ij}, husEduPri_{ij}, husEduSec_{ij}]$
- $\boldsymbol{x}_i = [rural_i, pcInd81_i]$
- τ^2 = between-family variation in baseline log odds
- ν^2 = between-community variation in baseline log odds

Model Structure

This figure has different indices: to be corrected in lecture



□ = known values

○ = unknown values

It will take a hot second to run

Model fitting

often issues with convergence

Convergence warning with defaults:

```
  Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glme
 Family: binomial  ( logit )
Formula: immun ~ kid2p + momEdPri + momEdSec + husEdPri + husEdSec + rural +
    pcInd81 + (1 | mom) + (1 | cluster)
   Data: dat

Random effects:
 Groups  Name        Variance Std.Dev.
 mom     (Intercept) 1.2357   1.1116
 cluster (Intercept) 0.5032   0.7094
Number of obs: 2159, groups:  mom, 1595; cluster, 161

(Intercept)  -0.7610     0.2800  -2.718  0.00657 **
kid2p         1.2812     0.1581   8.106 5.24e-16 ***
momEdPri      0.2793     0.1470   1.900  0.05747 .
momEdSec      0.2858     0.3248   0.880  0.37884
husEdPri      0.3762     0.1457   2.583  0.00979 **
husEdSec      0.3262     0.2733   1.194  0.23258
rural        -0.6691     0.2049  -3.266  0.00109 **
pcInd81      -0.9823     0.2486  -3.951 7.79e-05 ***

convergence code: 0
Model failed to converge with max|grad| = 0.0258663 (tol = 0.002, component 1)
```


STOPPED HEEEEEEEEEEEEEEEERRRRRRRRRRRREEEEEEEEEEE

Model Fit

```
> fit = glmer(immun ~ kid2p + momEdPri + momEdSec + husEdPri + husEdSec +
rural + pcInd81 + (1|mom) + (1|cluster), family = binomial, data = dat,
glmerControl(optimizer="bobyqa"))
> summary(fit)
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMo
 Family: binomial  ( logit )
Formula: immun ~ kid2p + momEdPri + momEdSec + husEdPri + husEdSec + rural +
    pcInd81 + (1 | mom) + (1 | cluster)
   Data: dat
Control: glmerControl(optimizer = "bobyqa")

Random effects:
 Groups  Name        Variance Std.Dev.
 mom     (Intercept) 1.2373   1.1123
 cluster (Intercept) 0.5038   0.7098
Number of obs: 2159, groups:  mom, 1595; cluster, 161

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.7624     0.2801  -2.722  0.00649 **
kid2p         1.2815     0.1581   8.105 5.27e-16 ***
momEdPri      0.2793     0.1471   1.899  0.05753 .
momEdSec      0.2808     0.3248   0.864  0.38737
husEdPri      0.3771     0.1457   2.588  0.00966 **
husEdSec      0.3308     0.2734   1.210  0.22629
rural        -0.6686     0.2049  -3.263  0.00110 **
pcInd81      -0.9824     0.2487  -3.950 7.83e-05 ***
---
```


Heterogeneity Interpretations

```
Random effects:
 Groups  Name        Variance Std.Dev.
 mom     (Intercept) 1.2373   1.1123
 cluster (Intercept) 0.5038   0.7098
Number of obs: 2159, groups:  mom, 1595; cluster, 161

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.7624     0.2801  -2.722  0.00650 **
```

- Baseline = rural community with 0% below poverty, mother and husband did not have primary or secondary education, and the child was under 2 years old. Baseline prob of immunization for a typical (population average) child in a model controlling for family & community intercepts:
$$\frac{e^{-0.7624}}{1+e^{-0.7624}} = 0.32$$
- Between-family variation contributes more than between-community variation.
- The total variation in baseline log odds has a standard deviation of $\sqrt{1.237+0.504} = 1.32$. 95% of the baseline probabilities are within
$$\frac{e^{-0.7624 \pm 1.96 \times 1.32}}{1+e^{-0.7624 \pm 1.96 \times 1.32}} = (0.03, 0.87).$$

Fixed-Effect Interpretations

Safe formula:

$100(e^{\beta}-1)$

```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.7624     0.2801  -2.722  0.00650 **
kid2p         1.2815     0.1581   8.105 5.27e-16 ***
momEdPri      0.2793     0.1471   1.899  0.05753 .
momEdSec      0.2808     0.3248   0.864  0.38735
husEdPri      0.3771     0.1457   2.588  0.00966 **
husEdSec      0.3308     0.2734   1.210  0.22629
rural        -0.6686     0.2049  -3.263  0.00110 **
pcInd81      -0.9824     0.2487  -3.950 7.83e-05 ***
```

After controlling for within-family and within-community correlation, or you could say "After ctrling for family & community-level intercepts"

- odds of immunization decreased in rural communities (OR: $e^{-0.67} = 0.51$, a 49% decrease), and communities with higher percentage of indigenous population: $1 - e^{-0.98*0.1} = 0.093$, a 9% decrease for 10% increase in indigenous population).
- higher immunization rate was associated with families where the mother (OR: $e^{0.28} = 1.31$) or the husband (OR: $e^{0.38} = 1.46$) received primary education vs without primary education. No significant effects for secondary edu (smaller sample size).
- children born during the campaign had a higher immunization rate (OR: $e^{1.28} = 3.59$). → this one we are interested in

$100(e^{-0.67}-1) = -49\%$ aka 49% decrease

$100(e^{-0.98 \cdot 0.1}-1) = -9\% = 9\%$ decrease

Regression Coefficients

What is the odds ratio for a child at least 2 years old (kid2p=1) versus less than 2 years old (kid2p=0) for a child from the same family and community, holding other variables constant?

$$e^{1.2815+u_i+u_{ij}+\boldsymbol{\beta}'\boldsymbol{x}_{ijk}} / e^{0+u_i+u_{ij}+\boldsymbol{\beta}'\boldsymbol{x}_{ijk}} = e^{1.2815}$$

$$e^{1.2815} = 3.60, \;\; 95\% \text{ CI} : e^{1.2815 \pm 1.96 * 0.1581} = [2.642, 4.911]$$

- The community and family random intercepts cancel out because we are holding them constant.
- However, the regression coefficients still have a conditional interpretation because the coefficients were estimated in the conditional model.

Working with Regression Coefficients

What is the odds ratio and 95% CI in immunization between two children from the same community and same mother with

- child B = over 2 years old at interview, mother's husband had primary education
  kid2p = 1
- child A = under 2 years old at interview, mother's husband had no primary education

$\hat{\beta}_{kid2p}$ $\hat{\beta}_{husband\ primary\ edu}$

$$\text{logit } (p_B) - \text{logit } (p_A) = 1.2815 + 0.3771$$

$$\text{OR for child B versus child A } = e^{1.2815+0.3771} = 5.25$$

Working with Regression Coefficients

```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
kid2p         1.2815     0.1581   8.105 5.27e-16 ***
husEdPri      0.3771     0.1457   2.588  0.00966 **
Correlation of Fixed Effects:
         (Intr) kid2p  mmEdPr mmEdSc hsEdPr hsEdSc rural
kid2p    -0.450
momEdPri -0.343  0.106
momEdSec -0.234  0.072  0.349
husEdPri -0.355  0.091 -0.154 -0.067
husEdSec -0.281  0.015 -0.172 -0.474  0.394
rural    -0.542 -0.075 -0.003  0.109  0.052  0.196
pcInd81  -0.438 -0.108  0.195  0.106  0.024  0.045  0.050
```

correlation

variance stuff

$$\begin{aligned} \text{Var(log OR)} &= 0.1581^2 + 0.1457^2 \\ &\quad + 2 \times \underbrace{0.091 \times 0.1581 \times 0.1457}_{\text{covariance}} = \\ \text{SE(log OR)} &= 0.225. \end{aligned}$$

$Var(\hat{\beta}_{kid2p} + \hat{\beta}_{huspri}) = Var(\hat{\beta}_{kid2p}) + Var(\hat{\beta}_{huspri}) + 2\,Cov(\ldots)$

So a 95% confidence interval is

$$\boxed{e^{(1.2815+0.3771) \pm 1.96 \times 0.225}} = (3.38,\ 8.16).$$

will do matrices when we got to splines