



Introduction to Multilevel Models

A PDHP workshop

Kristopher J. Preacher
Vanderbilt University

kris.preacher@vanderbilt.edu
<http://quantpsy.org>

Introduction

My goal is to cover some MLM basics, then tackle as many specific issues as I can.

Questions are great! Keep in mind that MLM is a very broad topic and we have only 4 hours to master it. Please limit questions to those that would interest 150+ other people.

If you want to know more about a particular topic, e-mail me; I will try to point you to relevant sources.



Introduction

A note on software:

I use R (lmer) and SPSS for most illustrations.

All examples (and more) are provided in Mplus code as well. Mplus is especially useful for (a) going beyond what most other MLM software can do (e.g., MSEM, multilevel mixture models, combining different variable types) and (b) power analysis.

An introduction to Mplus is included at the end of the slides for those interested.

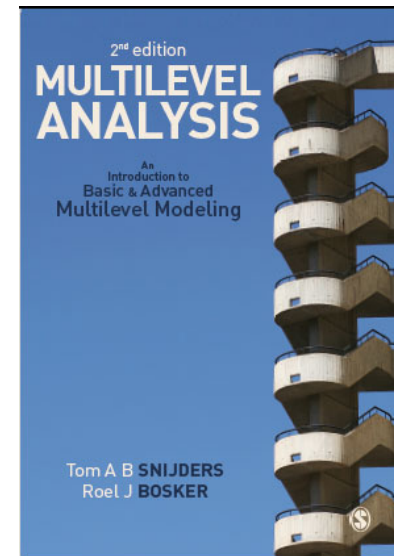
Books on multilevel modeling

The references at the end are much more complete. Here are the highlights:

- Hox, J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications* (3rd ed.). New York, NY: Routledge.
- Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.

I chose Snijders & Bosker (2012) as my semester course textbook. It has a good balance of coverage, price, and accessibility.

<http://www.stats.ox.ac.uk/~snijders/mlbook.htm>



Software options

Dedicated software packages (not exhaustive)

HLM
MLwiN

Other applications with multilevel capabilities (not exhaustive)

LISREL (MULTILEV)
EQS
SAS (PROC MIXED)
R (lmer, lme4, nlme, xxM)
S-Plus
SPSS (MIXED)
STATA
SYSTAT
Mplus

I chose SPSS for my semester course because its MIXED module is new and up to date, it is easy to use, it does everything I need in the course, and it is widely available.

<http://www-01.ibm.com/software/analytics/spss/>

For my research, I typically use Mplus and sometimes R (lmer).

Listserv

There is a popular multilevel modeling listserv I highly recommend checking out:

<http://www.jiscmail.ac.uk/lists/multilevel.html>

Outline of workshop

1. **Review of OLS regression**
2. **How not to deal with nested data**
3. **Some multilevel models**
4. **Model-building strategies**
5. **Effect size**
6. **Interactions**
7. **Centering**
8. **Power**
9. **Three-level models**
10. **A model for cross-classified data**
11. **Models for categorical outcomes**
12. **Introduction to Mplus**
13. **References**

1. Review of OLS regression

Review of OLS multiple regression

Before beginning our presentation of multilevel models, consider the following multiple linear regression (MLR) model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + e_i$$

Where the i subscript denotes individuals and k denotes the number of predictors.

Review of OLS multiple regression

Properties of OLS

$$\sum_{i=1}^N e_i = 0$$

$$\sum_{i=1}^N e_i^2 = \text{minimum possible}$$

$$\sum_{i=1}^N y_i = \sum_{i=1}^N \hat{y}_i$$

$$\text{cov}(x_i, e_i) = 0$$

Assumptions of MLR

MLR, like most statistical analyses, has assumptions that must be satisfied to ensure the quality of the results.

Usually, these are assumptions regarding the ability of the estimator to yield quality estimates.

Often these assumptions concern the population, not the sample, so we can 'check' the assumptions but cannot know with certainty whether the assumptions are met.

Assumptions of MLR

Key assumptions of MLR are:

- 1) Correct specification of relationship between IVs and DV
- 2) Inclusion of the important IVs in the model
- 3) Perfect reliability of IVs
- 4) Constant variance of errors (homoscedasticity)
- 5) Independence of errors
- 6) Normality of errors

Three of the regression assumptions apply to the *errors*. Errors are unobserved population quantities. Residuals are the corresponding observable sample quantities.

In many cases, we can examine the residuals to evaluate assumptions about the errors.

Gauss-Markov theorem

More formally, according to the Gauss-Markov Theorem...

$$\begin{array}{l|l} E(e_i) = 0 & \text{errors have sum and mean} = 0 \\ \text{var}(e_i) = \sigma_e^2 < \infty & \text{error variances are finite and homoscedastic} \\ \text{cov}(e_i, e_j) = 0 & \text{errors are uncorrelated} \end{array}$$

Other assumptions, added later:

$$e_i \stackrel{iid}{\sim} N(0, \sigma_e^2) \quad \text{cov}(e_i, x_i) = 0$$

(i.e., errors are assumed to be not only homoscedastic, but normally distributed, and errors and predictors are assumed uncorrelated.)

We also assume the x_i are "fixed" rather than "random" (more on this later).


Consequences of violating assumptions of OLS

Properties of good estimators ($\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators):

- unbiasedness $E(\hat{\beta}_1) = \beta_1$
- efficiency there is no other estimator with smaller SE
- consistency $SE(\hat{\beta}_1) \rightarrow 0$ as $N \rightarrow \infty$

When the assumptions in the Gauss-Markov Theorem are met, OLS estimates are the best possible.

Assumption violations can:

- introduce bias in point estimates
 - introduce bias in SEs (we often overestimate significance of coefficients)
 - compromise significance tests!
- 

Independence of errors

Independence of errors means there are no subgroups or clusters within the sample with correlated errors.

Nested (a.k.a., clustered, hierarchical) data can lead to violations of independence.

For example, nesting can occur when measuring students in the same classroom; clients seen by the same therapist; voters living in the same district; or repeated observations nested within the same person.

Violations of this assumption do not affect regression slope estimates, but do bias standard errors.

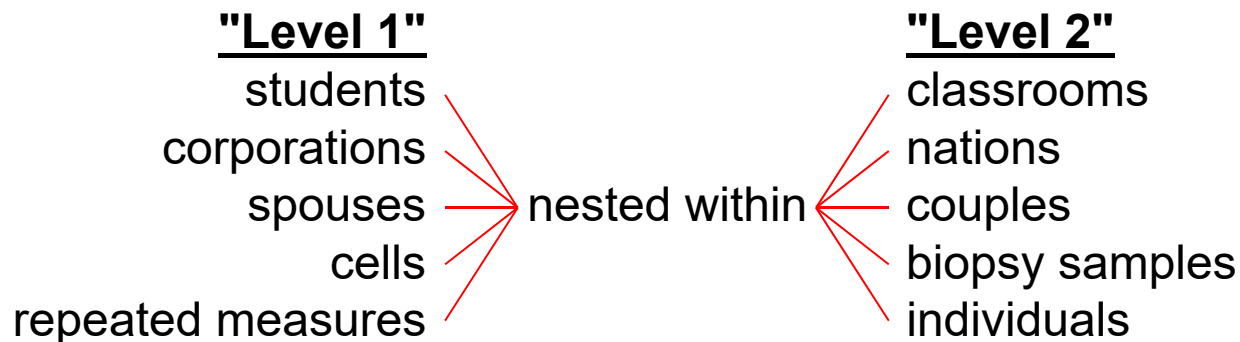
The severity of this violation depends on the degree of non-independence, but even small amounts of non-independence can have a noteworthy impact on Type I Error rates.

2. How not to deal with nested data

Dependence as a nuisance vs. interesting phenomenon

The assumptions of OLS are most likely to be met in simple random samples.

One problem is that, in practice, virtually no samples are simple random samples. Most involve clustering ("nesting") to some degree, sometimes to a large degree.



These are examples of 2-level *hierarchical data structures*.

Once you learn about hierarchical data, you start seeing examples everywhere (voters within precincts, patients within hospitals... etc.).

One consequence of nesting is that cases *within* clusters are typically more similar than cases *between* clusters, which potentially violates the assumptions of (a) homoscedasticity and (b) uncorrelated errors.

Dependence as a nuisance vs. interesting phenomenon


Ignoring violations of assumptions caused by nesting means we operate under the assumption that we have more information than we really do.

In clustered data, sampling units no longer yield unique information.

This leads to two primary problems:

1. Overestimating R^2

not necessarily, in light
of new research



2. Underestimating SEs of parameter estimates

Nesting can be a nuisance from a statistical point of view (no more Gauss-Markov = OLS is useless).

BUT, the dependence can also be a source of great substantive interest.

Methods of dealing with clustered data have moved gradually from *controlling for* clustering (attempting to remove its influence) to *modeling* it as something worth studying.

Addressing clustered data

What are some methods that have been used to address the multilevel nature of clustered data?

There are many methods, falling into three rough categories depending on how they consider "nestedness."

- Disaggregation
 - Aggregation
- } ignore nestedness
-
- ANCOVA
 - Separate regressions
 - Adjusting the results
 - Fixed effects approach
- } control or correct
for nestedness
-
- Multilevel modeling
- } *model* nestedness

Disaggregation

Also called pooled (or total) regression analysis.

Disaggregation involves analyzing the level-1 data as if there is no clustering. Level-1 and level-2 predictors (predictors measured at, e.g., the child and school levels) are included as if they were ordinary independent variables.

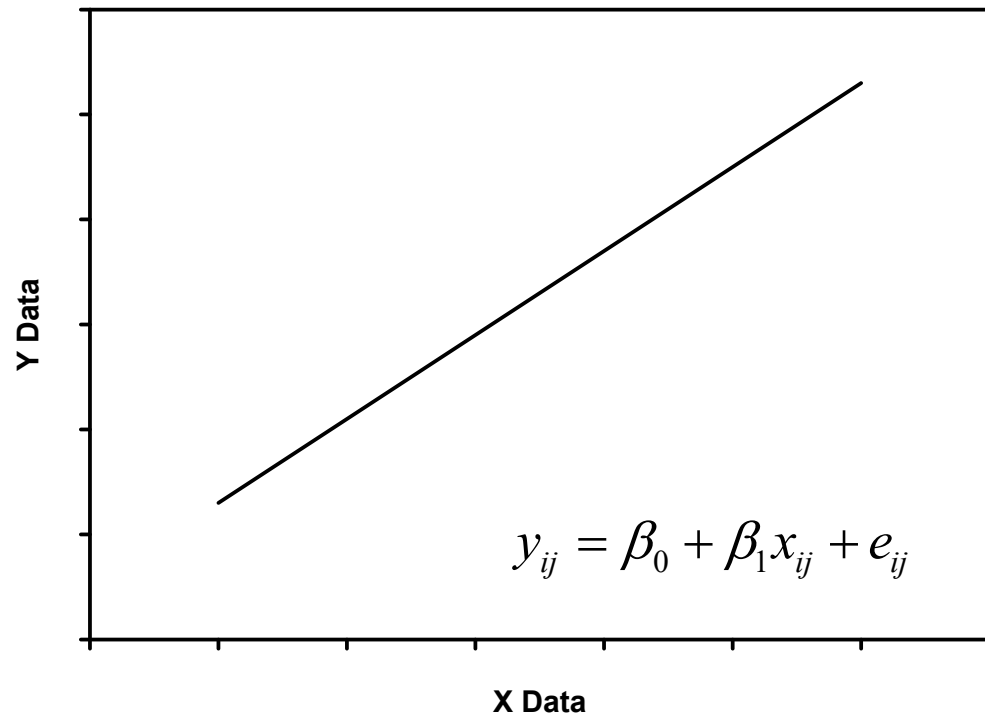
Level-1 units are the unit of analysis.

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 w_j + e_{ij},$$

In nested data, this can result in severe bias in standard errors due to violating the assumption of uncorrelated errors.

It also ignores between-cluster variation.

Not recommended.



Aggregation

Level-2 units are treated as the unit of analysis.

This procedure involves "aggregating to level-2," or averaging the values of each variable across members of each cluster.

$$y_{.j} = \beta_0 + \beta_1 x_{.j} + \beta_2 w_j + e_j,$$

Sample size is the number of level-2 units.

Clearly this procedure can result in a massive loss of information.

Interpretation is limited to level-2, which can be misleading (it may not be accurate to say the attitude of an organization toward some policy is simply the average of its employees' attitudes).

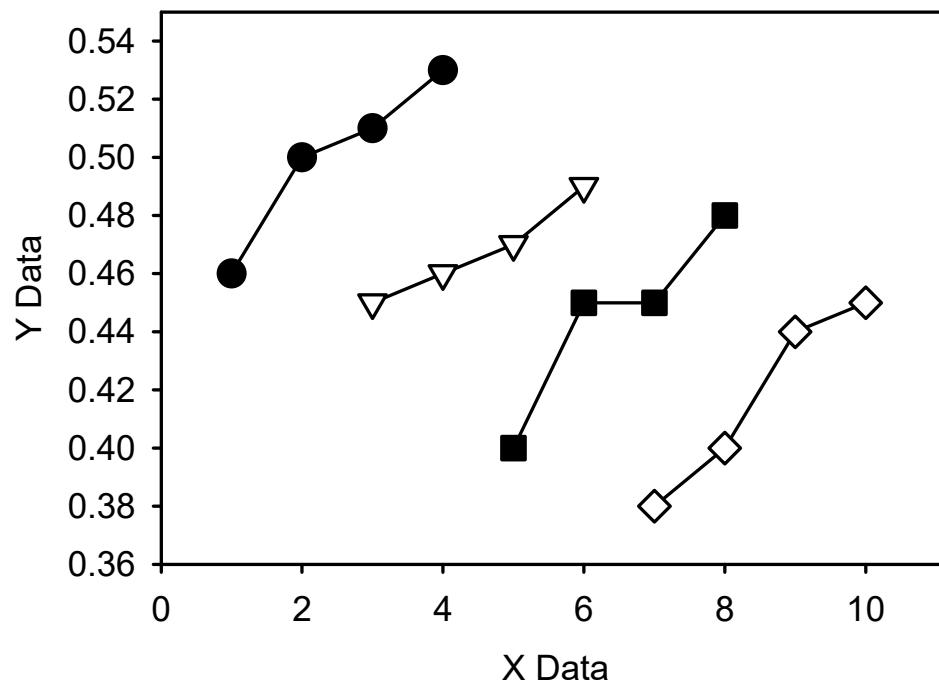
In psychology, we cannot deduce the behavior of crowds by weighting and averaging the behavior of individuals (e.g., the *bystander effect*).

Aggregation

Even more problems...

The results of aggregation analyses cannot be generalized to level-1 units (this is the *ecological fallacy*).

Anomalous effects, e.g.:



Regressions *within* clusters reveal positive trends, but if one averages *across* clusters, a negative trend emerges.

Not recommended...

Adjusting the results

This approach treats nestedness as a problem in need of solving.

Useful only if you are interested in aggregated effects, ignoring how effects may vary across clusters.

The logic: If nesting throws off standard errors in predictable ways, correct for it.

Example: Huber-White corrected standard errors, popular in econometrics as a way to adjust for non-independence of observations by correcting for correlated errors.

Not recommended.

Fixed-effects approach

This approach also "corrects for" nestedness, this time by including "group" as a predictor.

The researcher proceeds by including $J - 1$ dummy variables, such that $dummy_j = 1$ if a case belongs to group j , 0 otherwise.

$$y_i = \beta_0 + \sum_{j=1}^{J-1} \beta_j Dummy_i + e_i,$$

This method controls for *all* group differences.

However, the grouping variable is treated as "fixed," meaning that generalizability is restricted to only those groups represented in the sample.

In addition, there are many estimated parameters, so this model is not parsimonious.

Not recommended.

ANCOVA

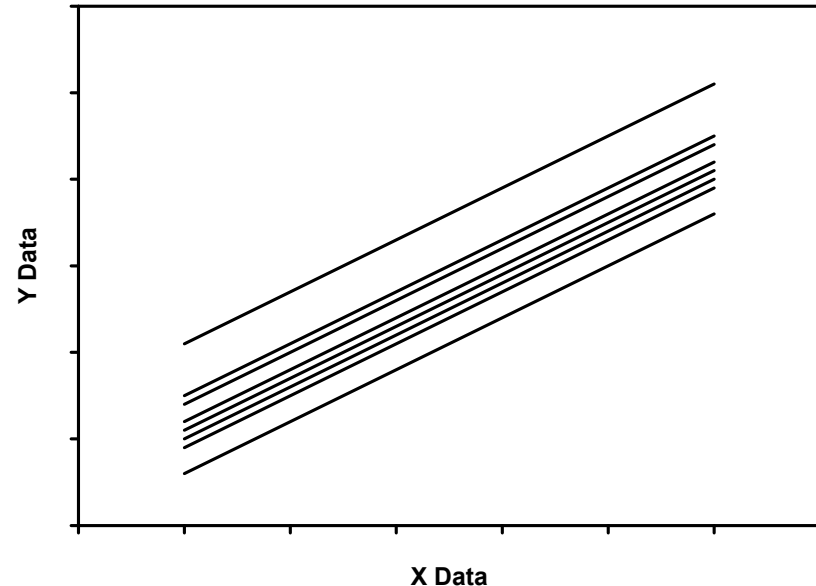
In analysis of covariance (ANCOVA), level-1 units are the unit of analysis.

The model is:

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + e_{ij},$$

The purpose is to test for an effect of level-2 units (clusters) on y , after removing the effect of level-1 covariates.

Intercepts can vary across clusters, but slopes cannot.



ANCOVA is useful for accommodating overall group effects, but...

- ...it does not permit varying slopes.
- ...it does not permit inclusion of level-2 predictors of intercepts.
- ...it does not impose a distribution for intercepts (intercepts are not *modeled*; they are simply *permitted to differ*.)

Not recommended.

Separate regressions

A more intuitive approach to treating nested data is to conduct a separate analysis within each level-2 cluster, yielding estimates of intercept and slope for each cluster.

We could examine the variance of intercepts, the variance of slopes, and their covariance.

We could examine the effects of level-1 predictors on y .

We could examine the effects of level-2 predictors on intercepts and slopes.

Limitations

- Impractical if there are many level-2 units.
- Estimates of slopes and intercepts are unreliable, but treated as error-free.
- Estimates of slopes and intercepts are given equal weight regardless of n_j .
- No partitioning of variance.
- Involves the estimation of a large number of parameters.

Not recommended.

3. Some multilevel models

Multilevel modeling

MLM treats clusters as if they are sampled from a larger population of clusters, enhancing the generalizability of results.

Cluster-level effects are not estimated separately for each cluster. Instead, regression weights are assumed to have a particular distribution across clusters, summarized by a limited set of parameters (e.g., mean and variance).

For example, whereas the single-level regression model might assume that all cases come from a population with the same intercept β_0 :

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad e_i \sim N(0, \sigma_e^2)$$

...MLM permits individual level-2 units to have their own distribution of β_0 . Rather than estimate each one individually, we assume a distributional form for β_0 as we do for e_i :

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij} \quad \begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} \sim MVN \left(\begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix} \right)$$
$$e_{ij} \sim N(0, \sigma_e^2)$$

Recommended !!

Random effects ANOVA / variance components analysis as MLM

A random effects ANOVA is like regular ANOVA, but "groups" are considered to be randomly sampled from a larger population.

$$y_{ij} = \beta_{0j} + e_{ij} \quad \text{"Level-1 model"} \quad e_{ij} \sim N(0, \sigma_e^2)$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad \text{"Level-2 model"} \quad u_{0j} \sim N(0, \tau_{00})$$

Reduced-form equation:

$$y_{ij} = \underbrace{\gamma_{00}}_{\text{fixed component}} + \underbrace{u_{0j} + e_{ij}}_{\text{random component}}$$

Properties:

$$\hat{\tau}_{00} + \hat{\sigma}_e^2 = \hat{\sigma}_y^2 \quad \text{(the between- and within-cluster variances sum to the observed variance, when computed using } N - 1.)$$

$$\text{ICC} = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}_e^2} \quad \text{(the intraclass correlation is the proportion of observed variance that is between units.)}$$

Intraclass correlation (ICC)

Some more on the ICC...

$$\text{ICC} = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}_e^2} \text{ is sometimes used to decide if MLM would be worthwhile.}$$

ICC is similar to R^2 in regression and the η^2 and ω^2 measures of effect size in ANOVA:

$$\eta^2 = \frac{SS_{\text{groups}}}{SS_{\text{total}}} \quad \omega^2 = \frac{SS_{\text{groups}} - (k - 1)MS_{\text{error}}}{SS_{\text{total}} + MS_{\text{error}}}$$

...and the proportion of explained variance in variance components analysis:

$$\%V = \frac{MS_{\text{groups}} - MS_{\text{error}}}{MS_{\text{groups}} + (n_j - 1)MS_{\text{error}}}$$

Example data set

Kanfer and Ackerman (1989) sampled 141 USAF personnel.

Each subject carried out a simulated air traffic control task in 3 – 6 trials.

In each trial, subjects were told to land several planes.

We are interested in the degree to which trials vary within-subject and between-subject. Is the observed variability mostly due to between-person differences or within-person differences?

kanfer.sav:

control = air traffic controller

time = number of the trial (1 – 6)

measure = number of successful landings per trial

ability = cognitive ability score

time_c = time centered at the first occasion

time_c2 = time centered and squared

★ I will note the existence of supporting code like this.

kanfer (R, SPSS, Mplus)

Kanfer & Ackerman air traffic control example

A random effects ANOVA model using the **air traffic control** example data set.

This is the simplest multilevel model, and can serve as a baseline to which to compare more complicated models, in which case it is termed a "null model."

The model is expressed as:

$$measure_{ij} = \beta_{0j} + e_{ij} \quad \leftarrow \text{level-1 model}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad \leftarrow \text{level-2 model}$$

Reduced form:

$$measure_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

Level-2 covariance structure:

$$u_{0j} \sim N(0, \tau_{00})$$

Level-1 covariance structure:

$$e_{ij} \sim N(0, \sigma_e^2)$$

Kanfer & Ackerman air traffic control example

Things to look for in random effects ANOVA...

Level-2 variance: $\hat{\tau}_{00} = 47.23(7.52)$ (highly significant)

Level-1 variance: $\hat{\sigma}_e^2 = 92.18(4.93)$ (highly significant)

$$\text{ICC} = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}_e^2} = \frac{47.23}{47.23 + 92.18} = 0.34$$

In other words, about 34% of the observed variation is due to differences among personnel.

Fixed vs. random coefficients

Fixed vs. random coefficients

A distinction of central interest; this is what differentiates multilevel regression from single-level regression.

Fixed: A coefficient is fixed if the same value is assumed to apply to all individuals.

Random: A coefficient is random if the values of the coefficient are assumed to have been drawn from a probability distribution.

Many coefficients in MLM have fixed and random “components.” The fixed component applies to all individuals in the sample, whereas the random component indicates departure of individuals from the fixed component.

In reality, the fixed component is merely the expected value (mean) of the distribution we assume for the random coefficient.

Fixed vs. random coefficients

The random effects ANOVA model again:

$$y_{ij} = \beta_{0j} + e_{ij} \quad \text{"Level-1 model"}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad \text{"Level-2 model"}$$

The *reduced form equation* for the random effects ANOVA:

$$y_{ij} = \underbrace{\gamma_{00}}_{\text{fixed component}} + \underbrace{u_{0j}}_{\text{random component}} + e_{ij}$$

random coefficient

β_{0j} is treated as a *random coefficient* as defined earlier, because (a) we are interested not in individual intercepts but rather in the mean (γ_{00}) and variance (τ_{00}) and (b) we want to generalize results beyond the particular groups in our study.

Some authors call (co)variances of coefficients (e.g., τ_{00}) “random coefficients,” but this is a misnomer; “coefficient” means “multiplicative factor” or “weight.”

Random effects ANCOVA

We can expand the random effects ANOVA model to include fixed level-1 predictors. The result is a *random effects ANOVA* model with covariates (i.e., RANCOVA):

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

The effect of x is constrained to be the same across all level-2 units. In other words, $\tau_{11} = 0$ (where τ_{11} is the slope variance). The intercept variance (τ_{00}), on the other hand, is allowed to be nonzero.

The reduced-form equation is:

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + u_{0j} + e_{ij}$$

The level-1 variance σ_e^2 is now the residual variance after adjusting for the predictor x .

The level-2 variance τ_{00} is the group-level residual variance.

Random effects ANCOVA

The only difference between RANCOVA and ANCOVA is that the “group effect” (u_{0j}) is treated as **random** rather than **fixed**.

In this context, that means we (a) generalize to a population of such groups and (b) estimate the mean and variance of the coefficient distribution.

We could expand the RANCOVA model to include more level-1 predictors or level-2 predictors. For example:


$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + \beta_{3j}x_{3ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}w_{1j} + \gamma_{02}w_{2j} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$\beta_{3j} = \gamma_{30}$$



This is technically not RANCOVA because it includes level-2 predictors.

Random effects ANCOVA

Example: Junior School Project (JSP) data (Mortimore, et al., 1988).

Math and language tests

Three consecutive years

1192 students from 49 London schools

Variables:

| | |
|----------|--|
| school | level-2 unit |
| student | level-1 unit |
| gender | (girls = 0, boys = 1) |
| ravens* | (visual-spatial fluid intelligence test) |
| math1 | |
| math2 | |
| math3* | |
| eng1 | |
| eng2 | |
| eng3 | |
| constant | 1.0 |

Random effects ANCOVA

Contents of file jsp.sps:

```
*JSP null model
MIXED MATH3
  /FIXED=INTERCEPT
  /METHOD=ML
  /PRINT=SOLUTION TESTCOV
  /RANDOM=INTERCEPT | SUBJECT (SCHOOL) COVTYPE (VC) .
```

Random effects ANCOVA

Contents of file jsp.sps:

```
*JSP level-1 predictor  
MIXED MATH3 WITH RAVENS  
  /FIXED=INTERCEPT RAVENS  
  /METHOD=ML  
  /PRINT=SOLUTION TESTCOV  
  /RANDOM=INTERCEPT | SUBJECT (SCHOOL) COVTYPE (VC) .
```


Multilevel regression models

Multilevel models extend single-level models by treating regression coefficients (intercepts, slopes) as dependent variables in their own right.

The dependent variable (y_{ij}) is always measured at the lowest level, Level 1.

Even though it is not possible to use a level-1 variable as a predictor in a level-3 equation, it is possible for the level-3 intercept variance to be reduced by x_{ij} . We can explain level-3 variance with level-1 predictors.

For example, differences among classes may be partially explained by considering a student-level variable.

Random effects can exist at any level *except the highest level*. For example, level-1 slopes can vary randomly across level-2 units, but in a two-level model there are no level-3 units for level-2 slopes to vary across.

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}w_{1j} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}w_{1j} + u_{1j}$$

Null / random intercept model

You've decided to examine your data with multilevel modeling. Where do you start?

It is always wise to begin with a “null model,” another name for a random-effects ANOVA model, or a random intercept model with no predictors.

$$y_{ij} = \beta_{0j} + e_{ij} \quad e_{ij} \sim N(0, \sigma_e^2)$$
$$\beta_{0j} = \gamma_{00} + u_{0j} \quad u_{0j} \sim N(0, \tau_{00})$$

This model permits us to examine how much variability exists at level-1 and level-2.

$$y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \quad \text{ICC} = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}_e^2}$$

But let's say you want to include a predictor of y_{ij} , also measured at level-1.

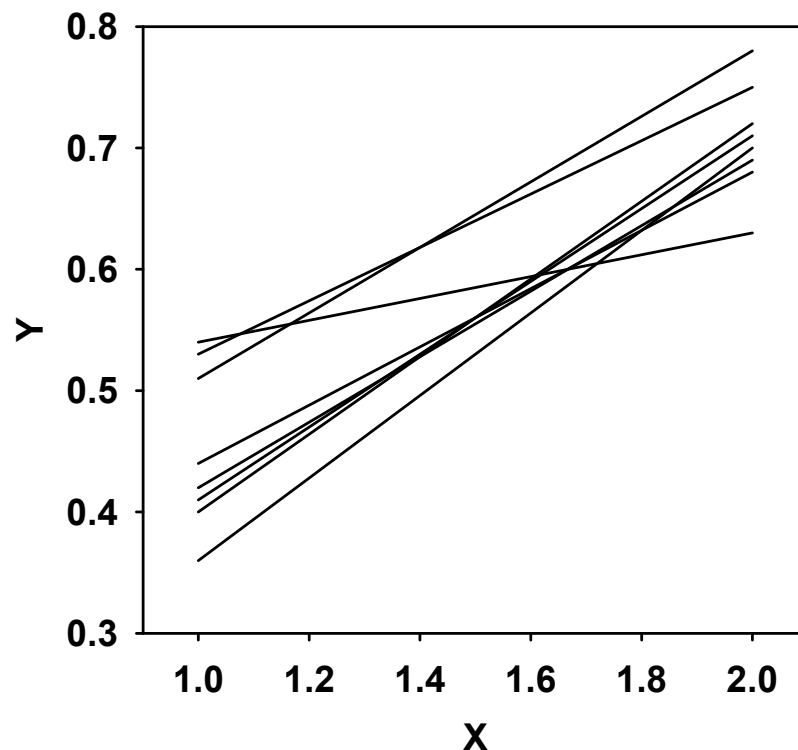
Separate regressions (review)

An intuitive “two-step” approach to treating nested data is to conduct a separate regression analysis within each level-2 cluster, yielding estimates of intercept and slope for each cluster.

We could examine the variance of intercepts, the variance of slopes, and their covariance.

We could examine the effects of level-1 predictors on y .

We could examine the effects of level-2 predictors on intercepts and slopes.



Separate regressions (review)

Limitations

- Impractical if there are many level-2 units.
- Estimates of slopes and intercepts are unreliable, but treated as error-free.
- Estimates of slopes and intercepts are given equal weight regardless of n_j .
- No partitioning of variance.
- Involves the estimation of a large number of parameters.

Not recommended!

We can visualize how multilevel regression improves upon—yet simplifies—separate regressions.

One level-1 predictor: random intercept, fixed slope

Let's say we wanted to fit a model with a random intercept and fixed slope. This is equivalent to the **random effects ANCOVA** model. Separate regressions and random coefficient regression will yield different results.

Random coefficient (multilevel) model equation:

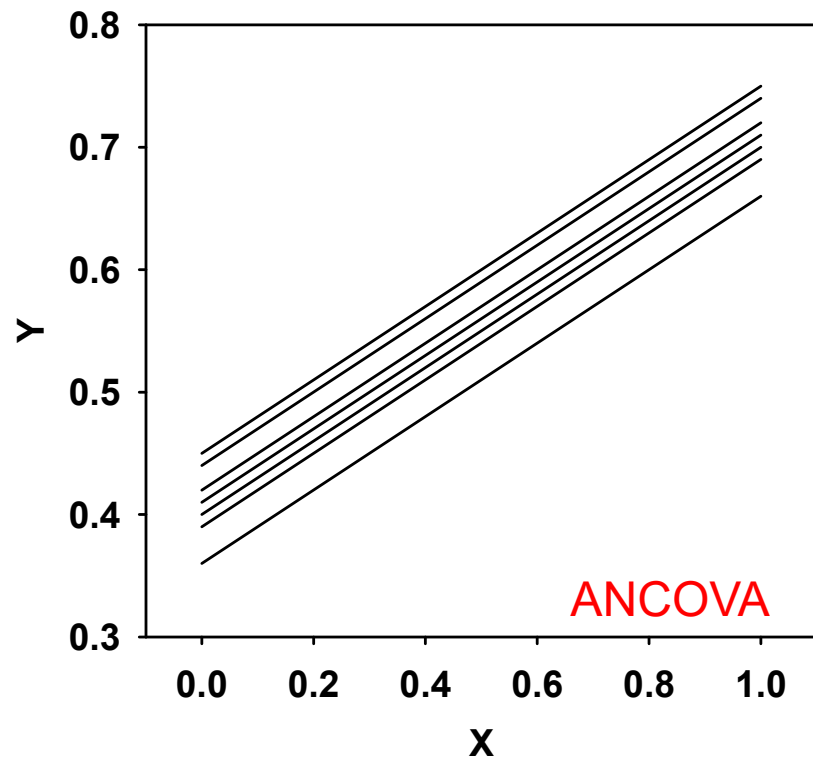
$$\begin{aligned}y_{ij} &= \beta_{0j} + \beta_{1j}x_{ij} + e_{ij} & e_{ij} &\sim N(0, \sigma_e^2) \\ \beta_{0j} &= \gamma_{00} + u_{0j} & u_{0j} &\sim N(0, \tau_{00}) \\ \beta_{1j} &= \gamma_{10}\end{aligned}$$

Reduced-form:

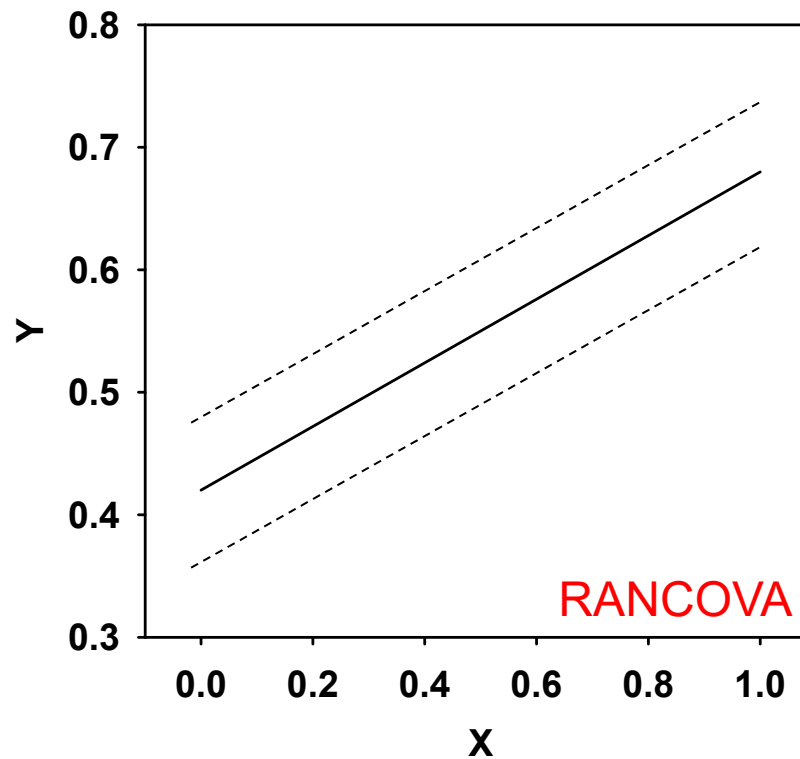
$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + u_{0j} + e_{ij}$$

One level-1 predictor: random intercept, fixed slope

Separate Regressions



Random Intercept / Fixed Slope
Multilevel Model



One level-1 predictor: fixed intercept, random slope

Let's say we wanted to fit a model with a fixed intercept and random slope. Separate regressions and random coefficient regression will yield different results.

Random coefficient (multilevel) model equation:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij} \quad e_{ij} \sim N(0, \sigma_e^2)$$

$$\beta_{0j} = \gamma_{00}$$

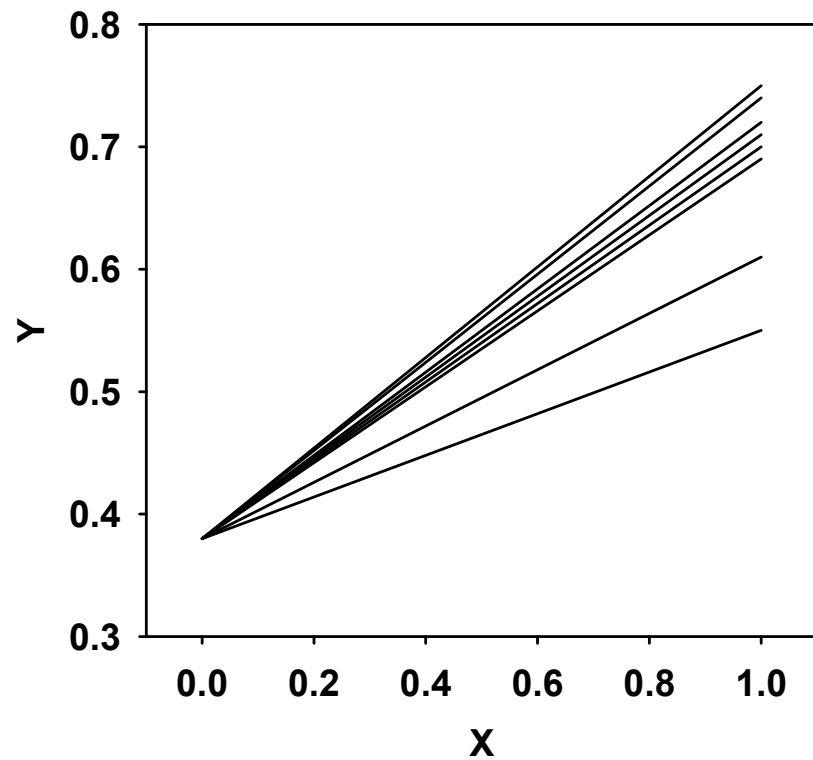
$$\beta_{1j} = \gamma_{10} + u_{1j} \quad u_{1j} \sim N(0, \tau_{11})$$

Reduced-form:

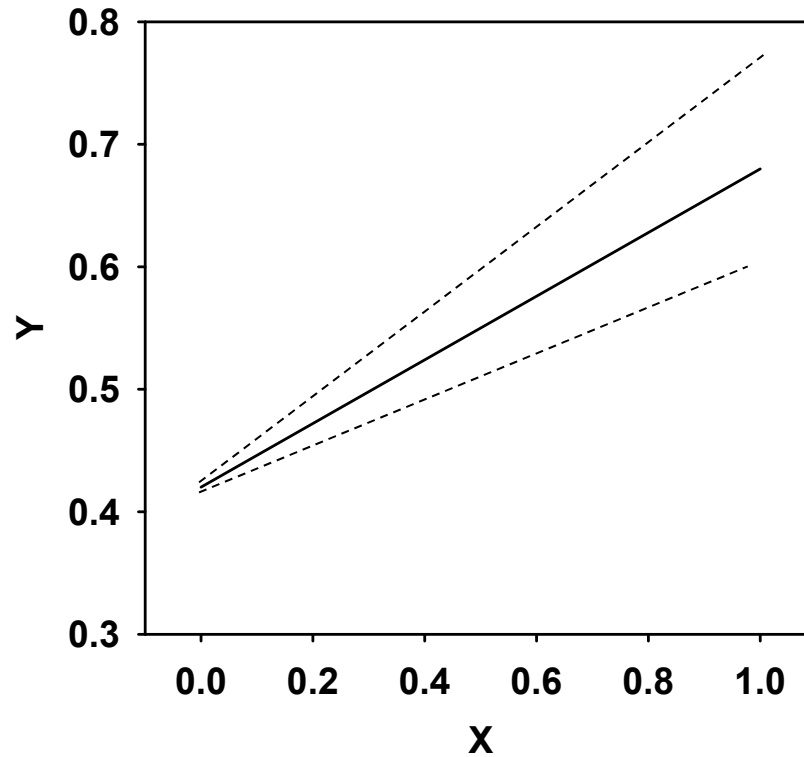
$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + u_{1j}x_{ij} + e_{ij}$$

One level-1 predictor: fixed intercept, random slope

Separate Regressions



Fixed Intercept / Random Slope
Multilevel Model



One level-1 predictor: random intercept, random slope

Let's say we wanted to fit a model with a random intercept and random slope. Separate regressions and random coefficient regression will yield different results.

Random coefficient (multilevel) model equation:

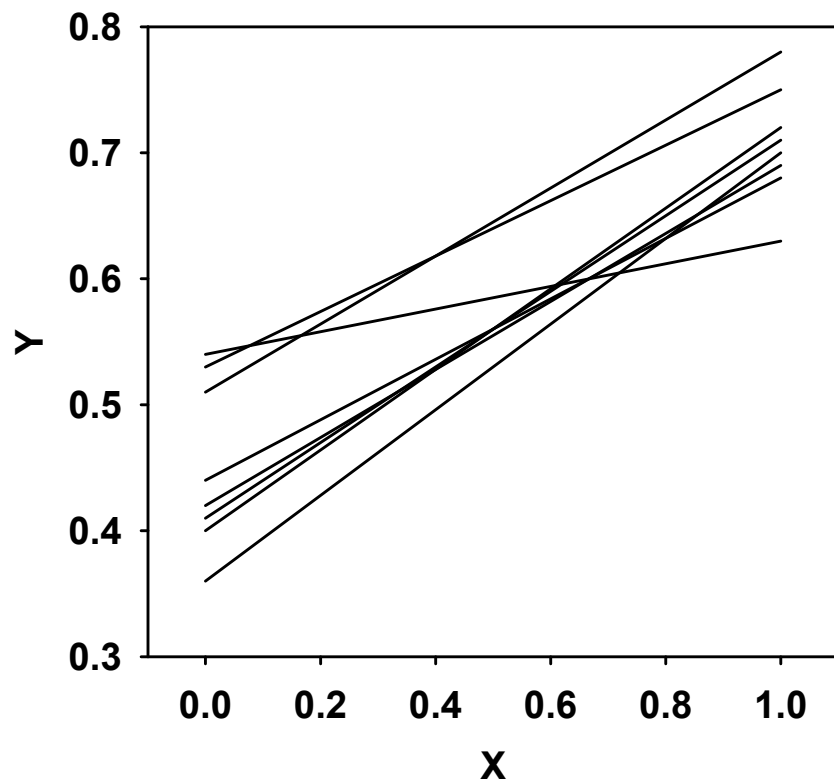
$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij} \quad e_{ij} \sim N(0, \sigma_e^2)$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$
$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix} \right)$$

Reduced-form:

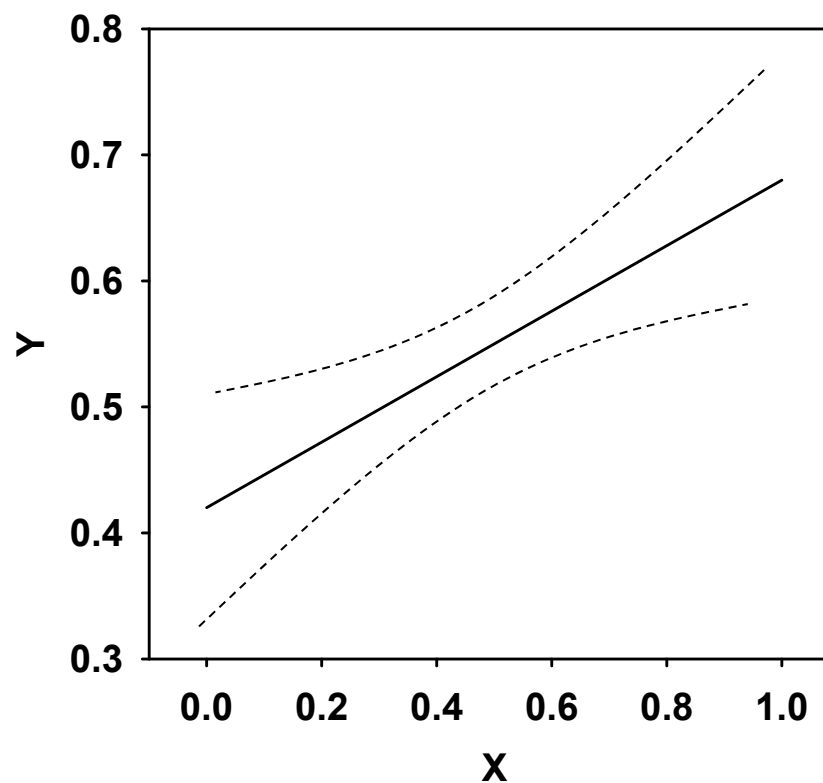
$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + u_{0j} + u_{1j}x_{ij} + e_{ij}$$

One level-1 predictor: random intercept, random slope

Separate Regressions



Random Intercept / Random Slope
Multilevel Model



Multiple level-1 predictors

In principle, we can have any number of level-1 predictors, and any mix of fixed and random coefficients. For example:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + \beta_{3j}x_{3ij} + \beta_{4j}x_{4ij} + \beta_{5j}x_{5ij} + \beta_{6j}x_{6ij} + \beta_{7j}x_{7ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + u_{2j}$$

$$\beta_{3j} = \gamma_{30} + u_{3j}$$

$$\beta_{4j} = \gamma_{40} + u_{4j}$$

$$\beta_{5j} = \gamma_{50} + u_{5j}$$

$$\beta_{6j} = \gamma_{60} + u_{6j}$$

$$\beta_{7j} = \gamma_{70}$$

$$e_{ij} \sim N(0, \sigma_e^2)$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \\ u_{5j} \\ u_{6j} \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & & & & & & \\ \tau_{10} & \tau_{11} & & & & & \\ \tau_{20} & \tau_{21} & \tau_{22} & & & & \\ \tau_{30} & \tau_{31} & \tau_{32} & \tau_{33} & & & \\ \tau_{40} & \tau_{41} & \tau_{42} & \tau_{43} & \tau_{44} & & \\ \tau_{50} & \tau_{51} & \tau_{52} & \tau_{53} & \tau_{54} & \tau_{55} & \\ \tau_{60} & \tau_{61} & \tau_{62} & \tau_{63} & \tau_{64} & \tau_{65} & \tau_{66} \end{bmatrix} \right)$$

All these slopes have the same* interpretation they would in single-level regression.

Nonrandomly varying slopes

There is a sort of “middle ground” between fixed and random coefficients.

It is possible to specify a model in which (e.g.) level-1 slopes are permitted to vary across clusters, but in a completely deterministic manner.

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}w_j + \underline{\hspace{1cm}}$$

We might specify such a model for reasons of efficiency and parsimony when, for example, τ_{11} is (or is expected to be) very small.

In this model, the variation in β_{1j} is completely explained by (perfectly correlated with) w_j .

Why is it called “nonrandomly varying”? Because whereas we permit the slopes to vary, we are claiming we’ve completely explained the variability in slopes; i.e., there is no *residual* variance at level-2 for that effect.

The full (two-level) multilevel regression model

Including level-2 predictors of slopes introduces interaction terms, whether we like it or not!

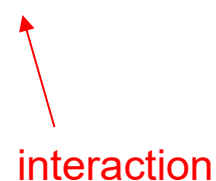
Level-1 and level-2 formulation:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij} \quad e_{ij} \sim N(0, \sigma_e^2)$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}w_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}w_j + u_{1j}$$
$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix}\right)$$

Reduced-form equation:

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}w_j + \gamma_{11}x_{ij}w_j + u_{0j} + u_{1j}x_{ij} + e_{ij}$$

conditional effects

interaction

The full (two-level) multilevel regression model

In older multilevel programs, specifying an interaction (cross-level or otherwise) required first computing the product term and entering it as a fixed-effect predictor.

SPSS allows specifying 2-way interactions directly in the syntax file:

```
/FIXED = INTERCEPT GENDER RAVENS GENDER*RAVENS
```

This method also works for 3-way (or higher-order) interactions.

Another way to think about it...

Here is a nifty way to think about MLM equations, especially in terms of translating a model into language that SPSS understands. Say this is your model:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + \beta_{3j}x_{3ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}w_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}w_j + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}w_j + u_{2j}$$

$$\beta_{3j} = \gamma_{30}$$

In reduced form...

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{1ij} + \gamma_{20}x_{2ij} + \gamma_{30}x_{3ij} + \gamma_{01}w_j + \gamma_{11}x_{1ij}w_j + \gamma_{21}x_{2ij}w_j \\ + u_{0j} + u_{1j}x_{1ij} + u_{2j}x_{2ij} + e_{ij}$$

Another way to think about it...

Before, we split this up into the fixed component and the random component:

$$y_{ij} = \underbrace{\gamma_{00} + \gamma_{10}x_{1ij} + \gamma_{20}x_{2ij} + \gamma_{30}x_{3ij} + \gamma_{01}w_j + \gamma_{11}x_{1ij}w_j + \gamma_{21}x_{2ij}w_j}_{\text{fixed}} + \underbrace{u_{0j} + u_{1j}x_{1ij} + u_{2j}x_{2ij}}_{\text{random}} + e_{ij}$$

In some situations it will be important to also provide structure to the (co)variances of the “e” residual terms as well (using the `/REPEATED` command), but not yet.

Recalling that we can add 1’s wherever we want...

$$y_{ij} = \underbrace{\gamma_{00}1 + \gamma_{10}x_{1ij} + \gamma_{20}x_{2ij} + \gamma_{30}x_{3ij} + \gamma_{01}w_j + \gamma_{11}x_{1ij}w_j + \gamma_{21}x_{2ij}w_j}_{\text{fixed}} + \underbrace{u_{0j}1 + u_{1j}x_{1ij} + u_{2j}x_{2ij}}_{\text{random}} + e_{ij}$$

In SPSS...

```
/FIXED = INTERCEPT x1 x2 x3 w x1*w x2*w  
/RANDOM = INTERCEPT x1 x2
```


4. Model-building strategies

Model-building strategies

When to use random coefficient models:

- If the level-2 units are understood as a random sample drawn from a population of such units.
- If the researcher wishes to test the effects of group-level (level-2) variables.

The goal: To either **develop** or **test** a parsimonious model that (a) describes the data satisfactorily and (b) is interpretable substantively.

Earlier we built a model by starting with a null model and adding a level-1 predictor, then slowly freeing intercepts and slopes.

In general, models can be extended by

- adding level-1 predictors (adds variables and parameters)
- adding random effects (adds only parameters, but sometimes many)
- adding predictors of those random coefficients (more variables and parameters)
- examining cross-level interactions

Model-building strategies

For example, we might progress through these stages, each time examining the parameter estimates, statistical significance, differences in model fit, and changes in explained variance.

$$y_{ij} = \beta_{0j} + e_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$\beta_{2j} = \gamma_{20}$$

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + u_{2j}$$

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}w_{1j} + \gamma_{02}w_{2j} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}w_{2j} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}w_{1j} + \gamma_{22}w_{2j} + u_{2j}$$

Using theory to guide model development (deductive)

If your theory suggests that certain effects should be fixed whereas others are expected to vary across level-2 units, then use that information when building your model.

This **deductive** approach is generally regarded as more scientific by philosophers of science; proceeds from theory to test.

A danger: It is often easy to argue *for* expecting random effects, and difficult to argue *against* expecting them.

Using theory to guide model development (deductive)

Some definitions (for this workshop, anyway):

Strong theory (with a capital “S”): A theory whose predictions can be translated into point null hypotheses in the context of a confirmatory quantitative model.

$$H_0: \mu = .75$$

(implied by theory; what we hope to find)

$$H_1: \mu = \text{anything else}$$

strong theory (with a little “s”): A theory whose broad predictions can be contrasted against point null hypotheses in the context of a quantitative model.

$$H_0: \mu = 0$$

$$H_1: \mu = \text{anything else}$$

(implied by theory; what we hope to find)

weak theory: A theory that provides only vague structure to hypotheses; generally fleshed out using exploratory analyses; requires confirmation on a new sample.

Using theory to guide model development (deductive)

Nobody ever has “Strong” theories.

Most theories tested with MLM are “strong” or “weak.”

For strong theories, a good strategy is to devise a series of plausible models *a priori*, evaluate them, and compare the results.

The most parsimonious model that provides an adequate description of the data is the “winner.”

Model selection can be used for this.

Using empirical evidence to guide model development (inductive)

If we have no strong theories, we can still use an exploratory procedure to select a model.

This approach builds theory from data. It is less scientific, but is widely used and accepted.

Typically the researcher will employ a mix of inductive and deductive strategies in building and testing a model.

Two general approaches:

- Build-up
- Tear-down

Stepwise “build-up” procedure

1. Null model (random intercepts only; no predictors)

Provides a baseline against which to compare later models.

$$y_{ij} = \beta_{0j} + e_{ij} \quad e_{ij} \sim N(0, \sigma_e^2)$$
$$\beta_{0j} = \gamma_{00} + u_{0j} \quad u_{0j} \sim N(0, \tau_{00})$$

Reduced-form:

$$y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

Compute ICC:

$$\text{ICC} = \frac{\tau_{00}}{\tau_{00} + \sigma_e^2}$$

Stepwise “build-up” procedure

2. Random intercept, level-1 predictor with a fixed slope

Add one or more fixed-coefficient level-1 predictors to the equation.

$$\begin{aligned}y_{ij} &= \beta_{0j} + \beta_{1j}x_{ij} + e_{ij} & e_{ij} &\sim N(0, \sigma_e^2) \\ \beta_{0j} &= \gamma_{00} + u_{0j} & u_{0j} &\sim N(0, \tau_{00}) \\ \beta_{1j} &= \gamma_{10}\end{aligned}$$

Reduced-form:

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + u_{0j} + e_{ij}$$

“main effect” of x



Stepwise “build-up” procedure

3. Random intercept, level-1 and level-2 predictors with fixed slopes

Add one or more level-2 predictor of intercepts.

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij} \quad e_{ij} \sim N(0, \sigma_e^2)$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}w_j + u_{0j} \quad u_{0j} \sim N(0, \tau_{00})$$
$$\beta_{1j} = \gamma_{10}$$

Reduced-form:

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}w_j + u_{0j} + e_{ij}$$

“main effects” of x and w

Stepwise “build-up” procedure

4. Look for random effects on a variable-by-variable basis

For example...

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij} \quad e_{ij} \sim N(0, \sigma_e^2)$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}w_j + u_{0j} \quad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix}\right)$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$

Reduced-form:

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}w_j + u_{0j} + u_{1j}x_{ij} + e_{ij}$$

Are these variances big enough to matter? If not, remove them.

“main effects” of x and w

A reasonable alternative is to look for random slopes before including level-2 predictors of intercepts.

Stepwise “build-up” procedure

4. Look for random effects on a variable-by-variable basis (continued)

It is quite possible for a variable to have no mean (“main”) effect, but to have a large random effect variance.

Therefore, we can include in this step variables that were previously omitted in Step 2.

Stepwise “build-up” procedure

5. Look for cross-level interactions


If some level-1 slopes were random (i.e., if their variances were large enough to be interesting) and if theory permits it, consider including level-2 predictors of slopes.

For example...

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij} \quad e_{ij} \sim N(0, \sigma_e^2)$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}w_j + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}w_j + u_{1j} \quad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix}\right)$$

Reduced-form:

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + \gamma_{01}w_j + \gamma_{11}x_{ij}w_j + u_{0j} + u_{1j}x_{ij} + e_{ij}$$


“main effects” of x and w interaction of x and w

Stepwise “build-up” procedure

The “build-up” procedure is exploratory in nature.

It could capitalize on chance characteristics of the sample.

Therefore, it is a good idea to estimate the obtained model on new data to see if it still “works.”



What you might use to judge whether to keep a variable or random effect in the model:

- Significance tests for parameters
- Improvement in model fit
- Explained variance

A combination of all three of the above is the wisest route. There are problems with each criterion:

| | |
|----------------------------|---|
| Significance tests: | Sample size and power issues; (co)variances non-normal |
| Improvement in fit: | No absolute fit in MLM; uncertain baseline for comparison |
| Explained variance: | Very tricky in MLM; conflicting methods in the literature |

Stepwise “tear-down” procedure

The opposite of the “build-up” procedure.

Involves starting with all effects random, and removing those that are nonsignificant and those that cause estimation errors.

Random effects should generally be removed from slopes before intercepts.

Can easily wind up with different final models using the “build-up” and “tear-down” procedures.

5. Effect size

Variance reduction and variance explanation

The literature on R^2 in MLM is sparse, inconsistent, and limited in a number of ways. Recently Rights and Sterba (2019) described these limitations and addressed them by developing an integrative framework of R^2 measures.

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, 24, 309-338.

They first note that R^2 can generically be defined as:

$$R^2 = \frac{\textit{explained variance}}{\textit{outcome variance}}$$

Further defining an R^2 in the context of MLM involves two considerations:

- 1) what *outcome variance* is of interest (**total**, **within-cluster**, or **between-cluster**)?
--denominator of R^2
- 2) which sources contribute to *explained variance*?
--numerator of R^2

Variance reduction and variance explanation

To define both the outcome variance and the explained variance, they derive an analytic partitioning of variance.

$$\begin{aligned} & \text{MLM-implied total outcome variance} \\ &= \text{var}(y_{ij}) \\ & \quad \vdots \\ &= \underbrace{\text{var}_{f_1} + \text{var}_v + \sigma^2}_{\text{within-cluster}} + \underbrace{\text{var}_{f_2} + \tau_{00}}_{\text{between-cluster}} \end{aligned}$$

Each term reflects variance attributable to one of five specific **sources**:

- var_{f_1} → level-1 predictors via *fixed slopes*
- var_v → level-1 predictors via *random slope variation*
- σ^2 → level-1 residuals
- var_{f_2} → level-2 predictors via *fixed slopes*
- τ_{00} → cluster-specific outcome means via *random intercept variation*

Variance reduction and variance explanation

This decomposition yields a set of measures differing in:

- (1) definition of outcome variance
- (2) source(s) of explained variance

Some examples:

-the proportion of *total* variance explained by *level-1 predictors via fixed slopes*

$$R_t^{2(f_1)} = \frac{\text{var}_{f_1}}{\text{var}_{f_1} + \text{var}_v + \sigma^2 + \text{var}_{f_2} + \tau_{00}}$$

-the proportion of *within-cluster* variance explained by *level-1 predictors via fixed slopes*

$$R_w^{2(f_1)} = \frac{\text{var}_{f_1}}{\text{var}_{f_1} + \text{var}_v + \sigma^2}$$

-the proportion of *total* variance explained by *all predictors via fixed slopes* (analogous to the traditional OLS R^2)

$$R_t^{2(f)} = \frac{\text{var}_{f_1} + \text{var}_{f_2}}{\text{var}_{f_1} + \text{var}_v + \sigma^2 + \text{var}_{f_2} + \tau_{00}}$$

Variance reduction and variance explanation

The full decomposition yields many possible ways of defining R^2 , and the set of possible measures can be visualized in a bar chart.

Total measures:

$$R_t^{2(f_1)}, R_t^{2(f_2)}, R_t^{2(v)}, R_t^{2(m)},$$

$$R_t^{2(f)} = R_t^{2(f_1)} + R_t^{2(f_2)},$$

$$R_t^{2(fv)} = R_t^{2(f_1)} + R_t^{2(f_2)} + R_t^{2(v)},$$

$$R_t^{2(fvm)} = R_t^{2(f_1)} + R_t^{2(f_2)} + R_t^{2(v)} + R_t^{2(m)}$$

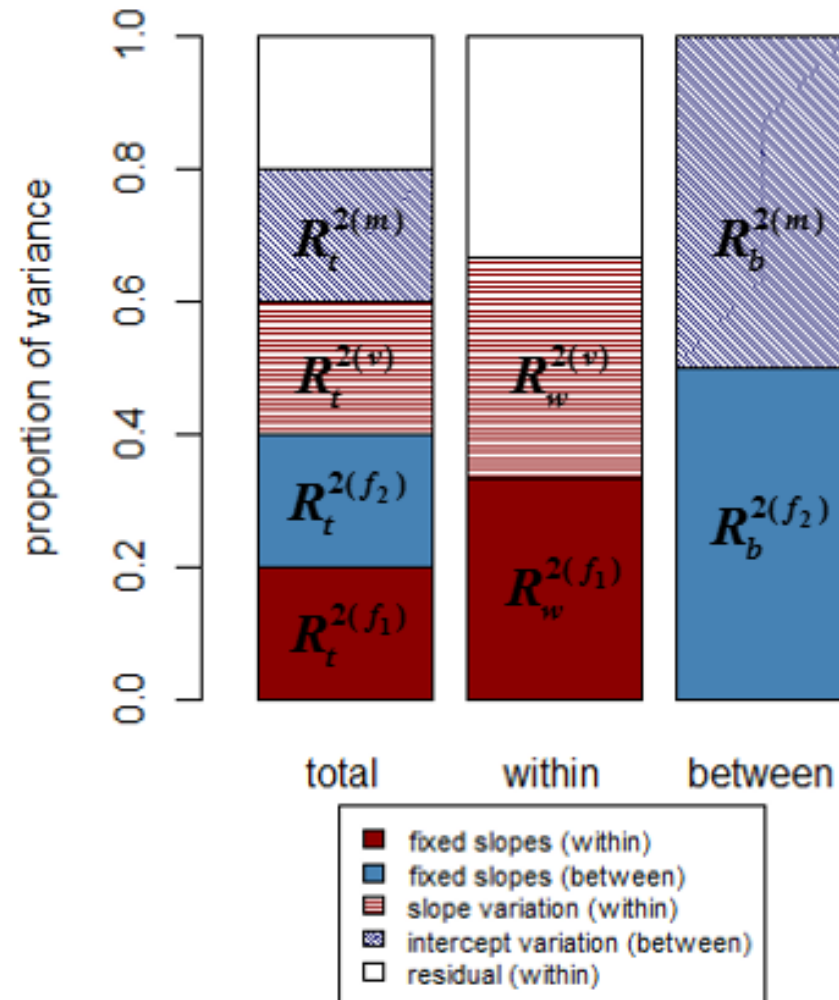
Within-cluster measures:

$$R_w^{2(f_1)}, R_w^{2(v)},$$

$$R_w^{2(f_1v)} = R_w^{2(f_1)} + R_w^{2(v)}$$

Between-cluster measures:

$$R_b^{2(f_2)}, R_b^{2(m)}$$



Variance reduction and variance explanation

This framework subsumes pre-existing measures, but also allows researchers to quantify variance explained in unique ways (e.g., quantifying variance explained by level-1 vs. level-2 predictors, quantifying variance explained by sources relative to total variance vs. level-specific variance, etc.).

| Author(s) | <i>Total measures</i> | | | | | | | <i>Within-cluster measures</i> | | | <i>Between-cluster measures</i> | |
|--|-----------------------|--------------|--------------|----------------|----------------|--------------|--------------|--------------------------------|----------------|--------------|---------------------------------|--------------|
| | $R_t^{2(f_m)}$ | $R_t^{2(f)}$ | $R_t^{2(f)}$ | $R_t^{2(f_1)}$ | $R_t^{2(f_2)}$ | $R_t^{2(v)}$ | $R_t^{2(m)}$ | $R_w^{2(f_1v)}$ | $R_w^{2(f_1)}$ | $R_w^{2(v)}$ | $R_b^{2(f_2)}$ | $R_b^{2(m)}$ |
| <i>Rights & Sterba (in press)</i> | | | | | | | | | | | | |
| <i>Vonesh & Chinchilli (1997)</i> | X | | X | | | | | X | | | | |
| <i>Snijders & Boskers (2012)</i> | | | X | | | | | | | | | |
| <i>Xu (2003)</i> | X | | | | | | | | | | | |
| <i>Aguinis & Culpepper (2015)</i> | | | | | | X | | | | | | |
| <i>Johnson (2014)</i> <i>(extension of Nakagawa and Schielzeth [2013])</i> | X | | X | | | | | | | | | |
| <i>Raudenbush & Bryk (2002)</i> <i>(first shown in 1992 edition;</i> <i>also in Kreft & DeLeeuw</i> <i>[1998] and Hox [2010])</i> | | | | | | | | X | | | X | |

Interactions: The example data

Data provided by Snijders and Bosker (1999).

2287 grade 8 pupils in 131 schools in The Netherlands (equivalent of U.S. 6th grade).

Data set contains information on IQ, grades, and different demographic and scholastic variables. We will use language test scores (LANGPOST), verbal IQ (IQ_VERB), and group size (GROUPSIZ).

6. Interactions

Interaction effects in OLS regression

Moderation: Occurs when the magnitude of the $x \rightarrow y$ relationship depends on z . If so, x *interacts with* z to predict y .

Moderation answers the question, "***When, under what circumstances, or for whom is x an important predictor of y ?***"

For example, maybe **verbal IQ** predicts **language test scores** more for children learning in small groups than for children learning in large groups (or vice versa).

In other words, **group size** may *moderate* the effect of **IQ** on **grades**.

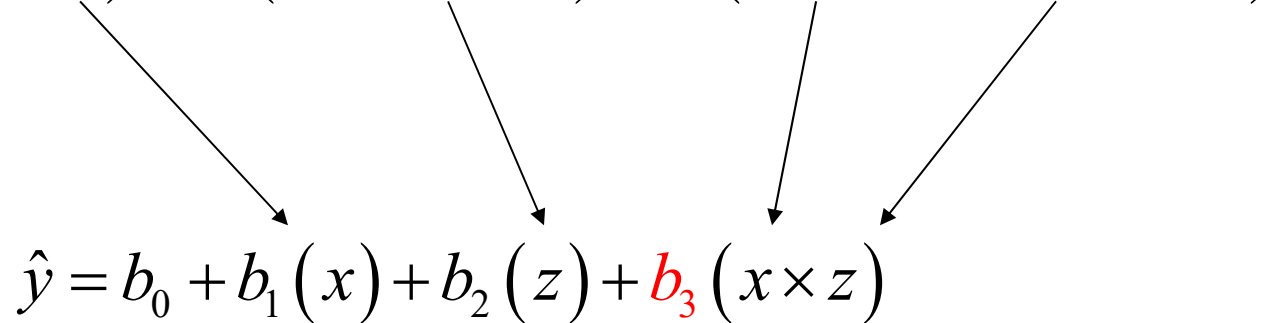
Focal predictors and moderators

Focal predictor (x): the independent variable of primary interest, whose relationship with y depends on the moderator.

Moderator (z): the independent variable that potentially predicts the $x \rightarrow y$ slope.

(I'll use b s instead of γ s for simplicity...)

$$\hat{y} = b_0 + b_1(\textit{focal}) + b_2(\textit{moderator}) + b_3(\textit{focal} \times \textit{moderator})$$


$$\hat{y} = b_0 + b_1(x) + b_2(z) + b_3(x \times z)$$

Okay, but what does it mean???

Simple intercepts and simple slopes

Return to the definition of moderation—when a slope can be expressed as a function of another variable.

Rearrange the prediction equation to be a linear function of the focal predictor:

$$\hat{y} = \underbrace{[b_0 + b_2z]}_{\text{simple intercept}} + \underbrace{[b_1 + b_3z]}_{\text{simple slope}} x$$

One could also obtain the simple slope with calculus as the derivative of y with respect to x , the (“instantaneous rate of change”).

Traditional method: Choose a few conditional values of the moderator and examine the simple slope of y on x for those values.

For continuous moderators, standard choices are the mean, -1SD , and $+1\text{SD}$.

Simple intercepts and simple slopes

Both the (simple) intercept and (simple) slope are **compound coefficients** that are **conditional on** the moderator.

$$\hat{y} = \underbrace{[b_0 + b_2z]}_{\text{simple intercept}} + \underbrace{[b_1 + b_3z]}_{\text{simple slope}} x$$

How to interpret the coefficients:

b_0 : The model-implied value of y when $x = z = 0$.
If predictors are centered, $b_0 =$ the mean of y .

b_1 : The expected x slope when $z = 0$.

b_2 : The expected z slope when $x = 0$.

b_3 : The expected change in the x slope per unit change in z .

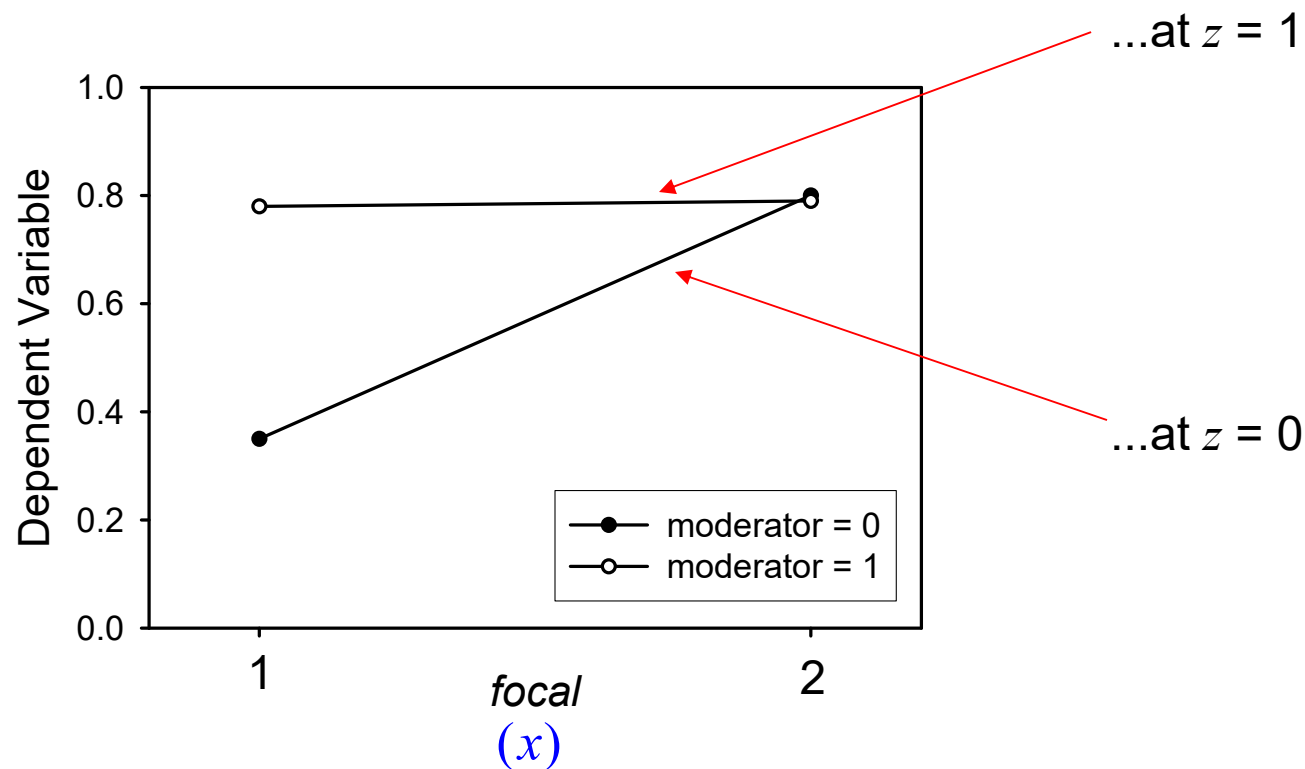
Plotting interaction effects

For dichotomous moderators...

Pick the coded values (usually 0 and 1) and plot the prediction equation for each value.

For example, a plot of

$$y = b_0 + b_1x + b_2z + b_3xz$$



Plotting interaction effects

In Snijders and Bosker's (1999) example...

$$\hat{y} = \underbrace{\left[b_0 + b_2 (\text{GROUPSIZ}) \right]}_{\text{simple intercept}} + \underbrace{\left[b_1 + b_3 (\text{GROUPSIZ}) \right]}_{\text{simple slope}} (\text{IQ_VERB})$$

For continuous moderators, standard choices are the mean, -1SD and +1SD, and/or Min and Max. Here, those are:

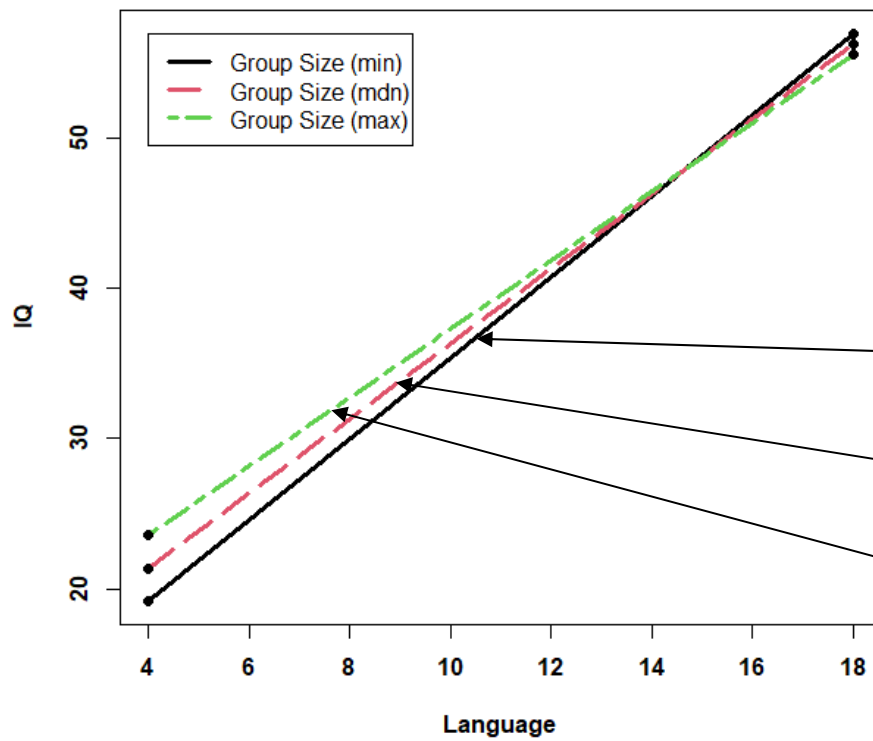
| | IQ_VERB | GROUPSIZ |
|------|----------------|-----------------|
| MIN | 4.000 | 5.000 |
| -1SD | 9.765 | 15.802 |
| MEAN | 11.834 | 23.101 |
| +1SD | 13.903 | 30.399 |
| MAX | 18.000 | 37.000 |

Plotting interaction effects

For Snijders and Bosker's (1999) example...

$$\hat{y} = \underbrace{\left[b_0 + b_2 (\text{GROUPSIZ}) \right]}_{\text{simple intercept}} + \underbrace{\left[b_1 + b_3 (\text{GROUPSIZ}) \right]}_{\text{simple slope}} (\text{IQ_VERB})$$

MLM 2-Way Interaction Plot



“Simple Regressions”

$$\hat{y} = \underbrace{\left[1.772 + .416(5) \right]}_{\text{simple intercept}} + \underbrace{\left[3.151 - .028(5) \right]}_{\text{simple slope}} (\text{IQ_VERB})$$

$$\hat{y} = \underbrace{\left[1.772 + .416(23) \right]}_{\text{simple intercept}} + \underbrace{\left[3.151 - .028(23) \right]}_{\text{simple slope}} (\text{IQ_VERB})$$

$$\hat{y} = \underbrace{\left[1.772 + .416(37) \right]}_{\text{simple intercept}} + \underbrace{\left[3.151 - .028(37) \right]}_{\text{simple slope}} (\text{IQ_VERB})$$

Probing interaction effects

Now that we have plotted the interaction, we may hypothesize that some simple slopes are different from zero and others are not. Investigating this is called “probing the interaction.”

“Probing” interactions means to explore the magnitude and significance of the relationship between y and x at different conditional values of z .

For example: “Is the effect of IQ_VERB on LANGPOST statistically significant for hypothetical individuals in groups of 30? How about for groups of 6?”

Probing interaction effects

Testing simple slopes is not as straightforward as testing, say, b_3 . However, it still involves dividing a point estimate by a standard error.

Assuming the simple slope is normally distributed across repeated sampling, we can use a t -test...

$$t = \frac{b_1 + b_3(z)}{\sqrt{s_{b_1}^2 + 2s_{b_1, b_3}(z) + s_{b_3}^2(z)^2}} \quad df = N - k$$

$(k = 4 \text{ here})$

The s terms come from the **asymptotic covariance matrix** (ACM or ACOV) of the parameter estimates, available* in any regression program. “ z ” is the conditional value of the moderator. “ b_1 ” and “ b_3 ” are estimated regression weights.

* (sometimes you have to really hunt for it; MLwiN, SAS, SPSS, R, Stata, LISREL, etc. all put it in different places)

The asymptotic covariance matrix

We need the (co)variances of the regression weights to test the simple intercept and simple slope for significance.

In SPSS, *some* of these are provided in standard regression output.

However, SPSS must be **tricked** into giving us the (co)variances associated with the intercept term.

This is another place where thinking of the intercept as a slope comes in handy...

$$ACM = \begin{bmatrix} \sigma_{b_0}^2 & & & \\ \sigma_{b_1b_0} & \sigma_{b_1}^2 & & \\ \sigma_{b_2b_0} & \sigma_{b_2b_1} & \sigma_{b_2}^2 & \\ \sigma_{b_3b_0} & \sigma_{b_3b_1} & \sigma_{b_3b_2} & \sigma_{b_3}^2 \end{bmatrix}$$

Testing simple slopes

In our example...

$$t = \frac{3.151 - .028(5)}{\sqrt{.064 + 2(-.00256)(5) + .000113(5)^2}} = \frac{3.009}{.2039} = 14.75$$

$$t = \frac{3.151 - .028(23)}{\sqrt{.064 + 2(-.00256)(23) + .000113(23)^2}} = \frac{2.498}{.0798} = 31.31$$

$$t = \frac{3.151 - .028(37)}{\sqrt{.064 + 2(-.00256)(37) + .000113(37)^2}} = \frac{2.101}{.1724} = 12.19$$

Simple slopes recap

- Run your regression analysis.
- Choose conditional values of the moderator.
- Find the simple slope (and intercept) associated with each conditional value of the moderator.
- Plot the conditional regression lines and determine the significance of their slopes
- Interpret.

Major limitation: *The conditional values are **arbitrary**.*

Johnson-Neyman technique / regions of significance

For what values of the moderator is the simple slope of y on x significant?

Start with the t formula...

$$t = \frac{b_1 + b_3(z)}{\sqrt{s_{b_1}^2 + 2s_{b_1, b_3}(z) + s_{b_3}^2(z)^2}}$$

Plug in the critical value for t ...


$$\pm 2.02 = \frac{b_1 + b_3(z)}{\sqrt{s_{b_1}^2 + 2s_{b_1, b_3}(z) + s_{b_3}^2(z)^2}}$$

...and solve for z (more difficult than it looks).

Johnson-Neyman technique / regions of significance

The result is a *region of significance*.

This region will either (1) *include* values of the moderator z that correspond to significant simple slopes and *exclude* those that do not:

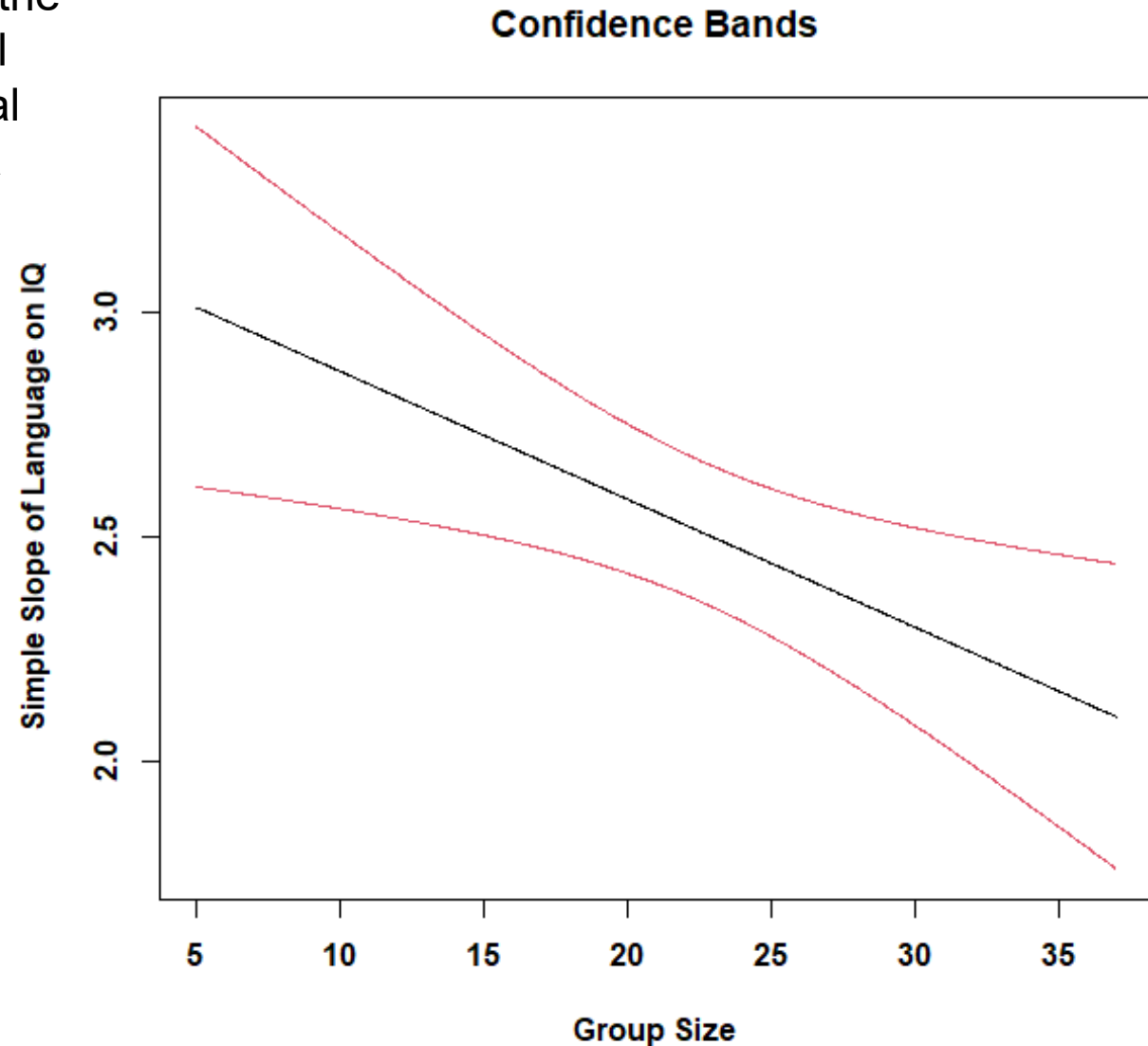
moderator range: 

...**or** (2) *exclude* values that correspond to significant simple slopes and *include* those that do not:

moderator range: 

Confidence bands

The horizontal axis here is the moderator (z). Each vertical slice is a confidence interval around a simple slope for a conditional value of z .



Plotting and probing cross-level interactions

Plotting and probing interactions in MLM is done in **exactly the same way** as in MLR, using exactly the same formulae for compound coefficients (point estimates of simple intercept and simple slope) and the standard errors.

The Johnson-Neyman technique for identifying regions of significance is done in precisely the same way, and confidence bands are produced in the same way.

The **only differences** lie in understanding the nature of multilevel modeling, and knowing where to fetch the ACM.

In both cases (MLR and MLM) only the fixed effects—and the asymptotic (co)variances of the relevant parameter estimates—are used.

Plotting and probing cross-level interactions

If using R...

To obtain the ACM:

- use `summary(snijders.intxn) $vcov`

If using SPSS...

To obtain the ACM:

- use `/PRINT = COVB`

If using Mplus...

To obtain the ACM:

- use `OUTPUT: TECH3;`

Use the HLM 2-way interaction online calculator to plot the interaction, compute regions of significance, and plot confidence bands.

Special considerations: Three-way interactions

With three interacting variables, there are several possibilities.

For example, three level-1 predictors:

$$y_{ij} = \underbrace{\gamma_{00} + \gamma_{01}x_{1ij} + \gamma_{02}x_{2ij} + \gamma_{03}x_{3ij} + \gamma_{04}x_{1ij}x_{2ij} + \gamma_{05}x_{1ij}x_{3ij} + \gamma_{06}x_{2ij}x_{3ij} + \gamma_{07}x_{1ij}x_{2ij}x_{3ij}}_{\text{fixed part}}$$

Plotting these requires choosing conditional values of two moderators and plotting lines for the outcome regressed on the focal predictor.

$$\hat{y}_{ij} = \underbrace{\left(\gamma_{00} + \gamma_{02}\dot{x}_{2ij} + \gamma_{03}\dot{x}_{3ij} + \gamma_{06}\dot{x}_{2ij}\dot{x}_{3ij}\right)}_{\text{simple intercept}} + \underbrace{\left(\gamma_{01} + \gamma_{04}\dot{x}_{2ij} + \gamma_{05}\dot{x}_{3ij} + \gamma_{07}\dot{x}_{2ij}\dot{x}_{3ij}\right)}_{\text{simple slope}} \dot{x}_{1ij}$$

All 3-way possibilities:

- 1 × 1 × 1
- 2 × 1 × 1
- 2 × 2 × 1
- 2 × 2 × 2

Special considerations: Dichotomous vs. continuous variables

Much has been said about the distinction between (and various combinations of) dichotomous and continuous predictors in interactions.

The methods are exactly the same. The only issue arises at the interpretation stage.

We always limit interpretation to observed values. In the special case of dichotomous variables, there are only two values.

If the **focal predictor** is dichotomous, then simple regressions represent the expected mean difference at various levels of the moderator.

If the **moderator** is dichotomous, then simple regressions represent the relationship between the outcome and the focal predictor at only two values of the moderator, and the other values must be ignored.

7. Centering

Centering

Centering refers to subtracting a value from raw data. This value is usually the **grand mean**, the **group mean**, or a **reference value** of particular interest.

Reasons why centering is used:

- to facilitate interpretation
- to avoid multicollinearity
- to separate effects into within- and between-cluster components

Centering is widely recommended in a variety of situations.

Centering can be useful in some circumstances, but often it can provide no benefit and give you more opportunity to make errors. Think carefully before centering.

Centering: Consequences for parameter estimation

Regression intercepts are “interpreted,” or best understood, at the point where all predictors = 0.

In regression models *with no* higher-order terms (squared terms, products, etc.), centering changes only the intercept, leaving slopes unaffected.

In regression models *with* higher-order terms, centering affects all terms *except* the highest-order terms.

Centering has extra consequences in MLM, where we are often interested in interpreting random intercepts. Where along x the intercept is located (“centered”) will influence how we interpret the intercept mean and variance (and any covariances involving intercepts).

Centering: OLS vs. MLM strategies

Centering strategies used in OLS:

- Uncentered data
- Grand mean centering
- Conditional value centering

Centering strategies used in MLM:

- Uncentered data
- Grand mean centering
- Conditional value centering
- Group mean centering
- Group mean centering, + mean

Uncentered data: Using the raw scale

Uncentered data: Using raw data without centering.

$$x_{ij}$$

This approach is sensible most of the time, especially if x is already in a meaningful metric.

Possible drawback: The intercept is interpreted as the predicted value of y when all $x = 0$. Zero may fall outside the observed range of x (income, number of friends), or may not be meaningful at all (height). How, then, to interpret the intercept?

Usually we do not care about the intercept, but occasionally it is useful.

Often (in the social sciences) we care more about what is going on when $x =$ the mean than when $x = 0$.

Uncentered data: Using the raw scale

An example model: Random intercept, random slope

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + u_{0j} + u_{1j}x_{ij} + e_{ij}$$

$$E[y_{ij}] = E[\gamma_{00} + \gamma_{10}x_{ij} + u_{0j} + u_{1j}x_{ij} + e_{ij}]$$

$$E[y_{ij}] = \gamma_{00} + \gamma_{10}\bar{x}_{..}$$

Grand mean centering

Grand mean centering: Subtracting the grand mean from all x scores.

$$x_{ij} - \bar{x}_{..}$$

Each level-2 intercept (β_{0j}) is now interpreted as the predicted y value for that level-2 unit *at the grand mean of x_{ij}* . Thus the (β_{0j}) are *adjusted group means*, interpreted at the grand mean of x_{ij} .

The average level-1 unit's x_{ij} score is now 0.

γ_{00} now represents the mean of the adjusted group means of y when $x_{ij} - \bar{x}_{..} = 0$.

τ_{00} is now the variance of the adjusted means at the average level of x_{ij} .

This model is statistically equivalent to the uncentered model.

Grand mean centering

An example model: Random intercept, random slope

$$y_{ij} = \beta_{0j} + \beta_{1j} (x_{ij} - \bar{x}_{..}) + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

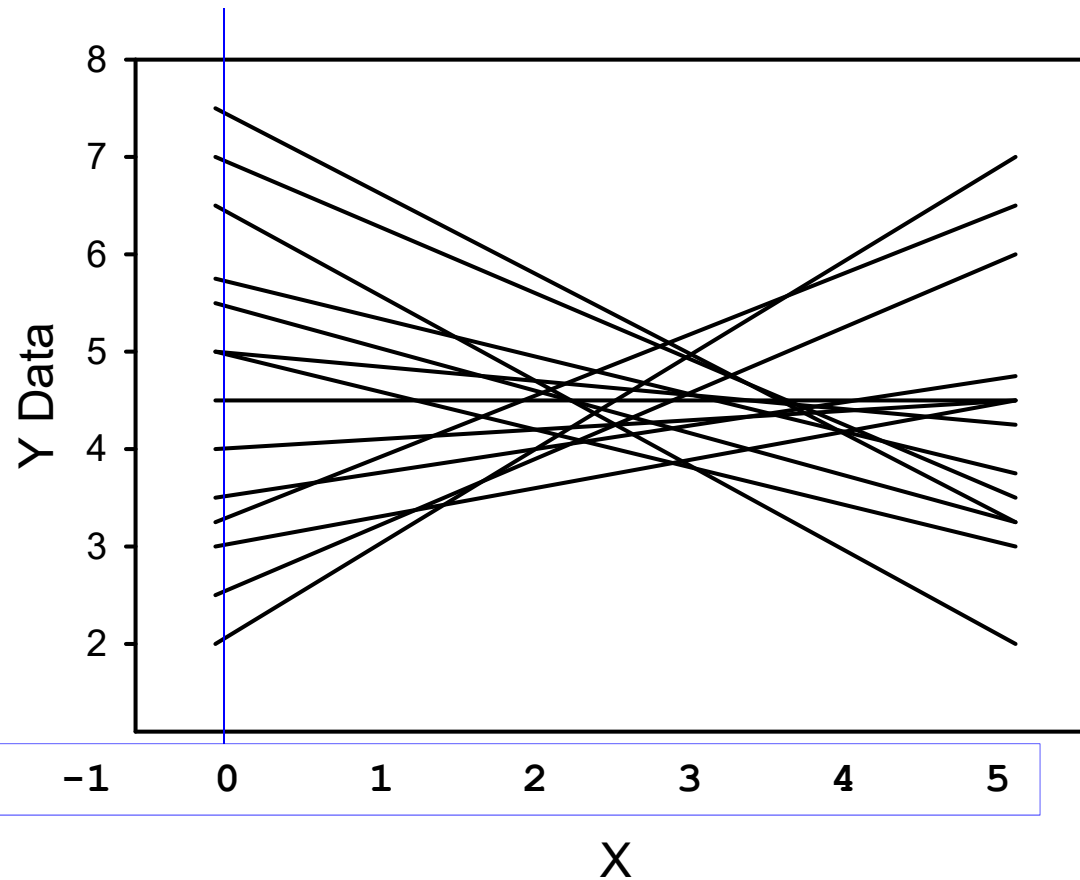
$$y_{ij} = \gamma_{00} + \gamma_{10} (x_{ij} - \bar{x}_{..}) + u_{1j} (x_{ij} - \bar{x}_{..}) + u_{0j} + e_{ij}$$

$$E[y_{ij}] = E[\gamma_{00} + \gamma_{10} (x_{ij} - \bar{x}_{..}) + u_{1j} (x_{ij} - \bar{x}_{..}) + u_{0j} + e_{ij}]$$

$$E[y_{ij}] = \gamma_{00}$$

i.e., γ_{00} represents the *expected value* of y_{ij} (the model-implied mean) under grand mean centering.

Grand mean centering: Consequences for intercepts



The intercept mean and variance (and any relevant covariances) change with centering.

Slopes do not change.

Cluster mean centering

Cluster mean centering: Subtracting the cluster mean from all x scores in each cluster.

$$x_{ij} - \bar{x}_{.j}$$

Because $\bar{x}_{.j}$ differs for each cluster, the zero-point will differ for each cluster. Thus, raw scores of 5.4 in one cluster and 3.6 in another might have the same “centered” value.

Each level-2 intercept (β_{0j}) is now interpreted as the predicted y value for that level-2 unit *at that cluster's mean value of x_{ij}* . Every regression line passes through $(\bar{x}_{.j}, \bar{y}_{.j})$, thus the (β_{0j}) are unadjusted cluster means.

Each level-2 unit's average x_{ij} score is now 0.

γ_{00} still represents the average level-1 unit's y_{ij} score.

Cluster mean centering

An example model: Random intercept, random slope

$$y_{ij} = \beta_{0j} + \beta_{1j} (x_{ij} - \bar{x}_{.j}) + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$y_{ij} = \gamma_{00} + \gamma_{10} (x_{ij} - \bar{x}_{.j}) + u_{1j} (x_{ij} - \bar{x}_{.j}) + u_{0j} + e_{ij}$$

$$E[y_{ij}] = E[\gamma_{00} + \gamma_{10} (x_{ij} - \bar{x}_{.j}) + u_{1j} (x_{ij} - \bar{x}_{.j}) + u_{0j} + e_{ij}]$$

$$E[y_{ij}] = \gamma_{00} \leftarrow \text{i.e., } \gamma_{00} \text{ represents the expected value of } y_{ij} \text{ under cluster mean centering.}$$

Cluster mean centering

Because cluster differences in x_{ij} are being subtracted (all cluster means now = zero), any effect observed for x_{ij} must reduce *only* level-1 variance, not level-2 variance.

Without centering, adding x_{ij} as a predictor could reduce level-1 and/or level-2 variance.

In that sense, the effect of a cluster mean centered variable is “cleaner.”

Adding the mean back in

We know from earlier that level-1 predictors x_{ij} can explain level-1 and/or level-2 variability. Cluster mean centering and using cluster means as level-2 predictors lets us separate these two effects, but we must “spend” one more parameter to do it.

Demonstration (due to Don Hedeker) that using raw score (or grand mean centered) x_{ij} confounds “within” and “between” explanation of variance:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij} \quad \leftarrow \text{level-1 equation}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad \leftarrow \text{level-2 equations}$$

$$\beta_{1j} = \gamma_{10}$$

$$y_{ij} = \gamma_{00} + u_{0j} + \gamma_{10}x_{ij} + e_{ij} \quad \leftarrow \text{reduced form expression}$$

$$y_{ij} = \gamma_{00} + u_{0j} + \gamma_{10} \left(x_{ij} - \bar{x}_{.j} + \bar{x}_{.j} \right) + e_{ij} \quad \leftarrow \text{playing with algebra}$$

$$y_{ij} = \gamma_{00} + u_{0j} + \gamma_{10} \left(x_{ij} - \bar{x}_{.j} \right) + \gamma_{10}\bar{x}_{.j} + e_{ij}$$

In other words, γ_{10} reflects both a within-cluster and between-cluster effect. This suggests a simple solution to the problem...

Adding the mean back in

This is not as simple as subtracting the mean in one step and adding it back in another. We *are* doing that, but estimating an extra parameter in the process.

$$y_{ij} = \gamma_{00} + u_{0j} + \gamma_{10} (x_{ij} - \bar{x}_{.j}) + \gamma_{01} \bar{x}_{.j} + e_{ij}$$

A nice feature: cluster means are uncorrelated with the cluster mean centered predictor, so $(x_{ij} - \bar{x}_{.j})$ decreases only level-1 variance and $\bar{x}_{.j}$ decreases only level-2 variance.

Cluster mean centering, unlike grand mean centering, is *not* merely a rescaling of x_{ij} to facilitate interpretation. It is a different model testing different hypotheses.

Consequently, all the parameter estimates have the potential to differ from corresponding parameters in uncentered or grand mean centered models.

Adding the mean back in

There are different ways to add cluster means to the model. For fixed slope models, one can use either x_{ij} and $\bar{x}_{.j}$ as predictors, or $(x_{ij} - \bar{x}_{.j})$ and $\bar{x}_{.j}$. These models are statistically equivalent.

However, if random slopes are involved, the two will no longer be equivalent because the latter will have an extra random effect term involving $\bar{x}_{.j}$.

Adding the mean back in

One use for cluster mean centering and re-including the cluster means is useful is estimation of the **contextual (compositional) effect**.

This effect is defined as the difference between the within-cluster and between-cluster effects.

The contextual effect is an idea commonly encountered in organizational research.

Conditional value centering

We need not always “center” using the mean.

Sometimes the median or mode are more sensible. Other times, specific conditional values (“reference values”) of x_{ij} are of interest.

In general, any parameters involving intercepts will be altered, but slopes will remain the same (unless interaction terms are included).

Always directly interpretable only where $x_{ij} = 0$.

Conditional value centering

Some special cases:

Sometimes the population mean of x_{ij} is known, and the researcher may want to generalize intercept-related results to the average population x_{ij} .

In longitudinal studies where repeated measures (level-1) are nested within people (level-2), we usually want to “anchor” the data at the initial occasion of measurement.

Interactions

Dummy variables...

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

For $x_{ij} = 0...$

$$y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

For $x_{ij} = 1...$

$$y_{ij} = \gamma_{00} + \gamma_{10} + u_{0j} + e_{ij}$$

We can switch the (0,1) coding and interpret γ_{00} as the mean of the other group.

Conditional value centering

Conditional value centering is sometimes used in the context of testing interactions.

Say you want to know the simple slope of y_{ij} on x_{ij} at a particular important value of w_j . (this works with any kind of interaction).

An easy way to “trick” software into helping probe interaction effects is to center w_j at the conditional value of the moderator. The reduced expression of a cross-level interaction:

$$y_{ij} = (\gamma_{00} + \gamma_{01}w_j) + (\gamma_{10} + \gamma_{11}w_j)x_{1ij} + u_{0j} + u_{1j}x_{1ij} + e_{ij}$$

When $w_j = 0$...

$$y_{ij} = (\gamma_{00}) + (\gamma_{10})x_{1ij} + u_{0j} + u_{1j}x_{1ij} + e_{ij}$$

$$y_{ij} = \gamma_{00} + \gamma_{10}x_{1ij} + u_{0j} + u_{1j}x_{1ij} + e_{ij}$$

simple intercept

simple slope

Both are now easy to test!
Requires data management.

Deciding how / whether to center

There is no universally correct choice for how (or whether) to center.

Do we prefer to make predictions about y_{ij} based on...

- Absolute level of x_{ij} ?
 - Use raw x_{ij} .
- Standing on x_{ij} relative to all other level-1 units?
 - Use grand mean centered x_{ij} .
- Standing on x_{ij} relative to all other level-1 units within a group?
 - Use group mean centered x_{ij} .
- Disentangling level-1 and level-2 effects?
 - Enter the group mean of x_{ij} as a level-2 predictor.

Does x have a meaningful zero? If not, then centering may be advisable.

How to cluster mean center

In SPSS...

Centering must be done manually.

Use the “aggregate” menu option.

Create a new variable = difference between original variable and cluster means.

In R (lmer)...

Centering must be done manually.

Compute grand means or cluster means.

Create a new variable = difference between original variable and cluster means.

In Mplus...

There are special commands for centering, e.g.:

DEFINE :

CENTER var1 var2 (GROUPMEAN) ;

Centering level-2 variables

At level 2, group mean centering is not an option (there are no higher-level groups).

The decision to grand mean center level-2 variables depends only on convenience and interpretability.

If w_j is already on a meaningful metric, there may be little reason to center.

If 0 is not meaningful or lies outside the range of interpretation, it may be sensible to center.

8. Power

Power—what is it?

Power: The probability of correctly rejecting a false null hypothesis, given that a particular alternative hypothesis is true.

Power is defined only when the null hypothesis is false. Realistically, ~100% of the time.

We want power to be high.

Power increases with...

- higher α (Type I error rate)
- larger sample size
- smaller error variance
- larger effect size

Of these, sample size is most directly under our control. Researchers usually want to know either:

1. how much power they can obtain with a given N .
2. how large an N is necessary to achieve a given level of power.

Power—what is it?

There is a large literature on power and sample size in MLM.

The best advice: **If you can get more data, do so**. MLM is a “large-sample” technique, meaning that assumptions are met to a greater degree with more confidence as N increases. So, the larger the sample, the better.

How large is “large”?

Singer & Willett (2003): “10 is certainly small and 100,000 is certainly large.”

Power—what is it?

Setting aside MLM for the moment, there are in general three kinds of power in SEM:

1. Power to reject a false model (“false” is defined as having sufficiently poor fit).
2. Power to reject the hypothesis that a parameter = 0.
3. Power for nested model tests (chi-square difference).

MLM does not have model fit indices, even for the class of MLM models that are equivalent to SEM.

We *can* speak of power to reject the null hypothesis that two nested models fit equally well, however (via the deviance test).

Generally power is considered for tests of **individual parameters**, usually **fixed effects**.

Power—what is it?

In addition to the factors mentioned earlier, power also depends on the method used to test the null, for example:

- Wald test

- Deviance test

- Likelihood-based confidence intervals

- Bootstrapping

...and on estimation method (FIML vs. REML).

Rules of thumb

In general, “more is better,” but more *what?*

There are (at least) two sample sizes to worry about in MLM, and the “balancedness” of level-1 n_j within level-2 units is also an issue, but a minor one.

For accurate standard errors for level-2 variances, a large level-2 sample size (J) is required.

Power analysis methods

There are a few different ways to think about, and conduct, power analysis in MLM.

I will cut to the chase and assert that the best method is the one based on *Monte Carlo simulation*. There are at least two ways to do this:

R:

- MLPowSim generates R code, which can then be modified and run.
- It must be modified because the generated R code has at least one deprecated command.
- It is limited in the types of models for which power analysis is possible.

Mplus:

- Uses Monte Carlo simulation method to estimate power.
- Virtually any model that Mplus can run on real data can also be used in power analysis.

Power analysis in Mplus: Monte Carlo simulation

Mplus' Monte Carlo facility has many uses. One is to conduct power analysis for individual parameters.

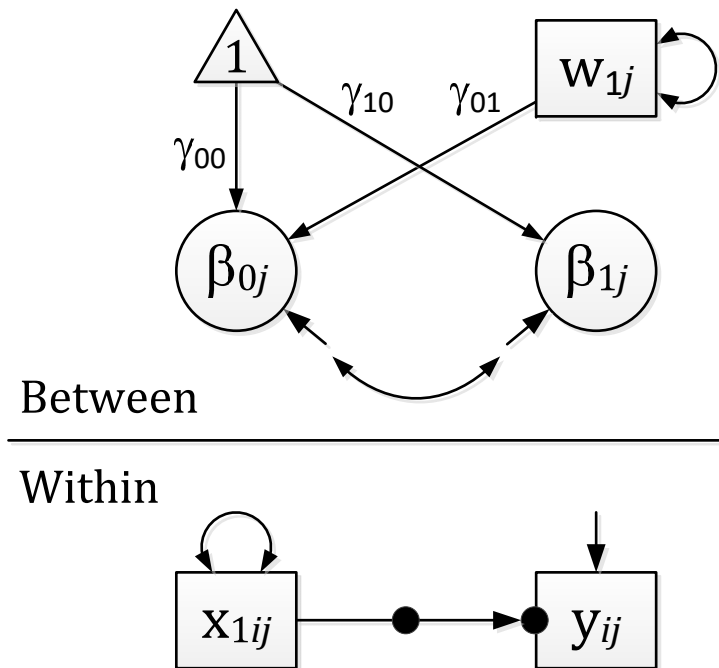
The basic idea:

- Specify a population model, with values given to all parameters.
- Decide on level-1 and level-2 sample sizes (J and n_j).
- Generate a large number of samples.
- Run the model on each of them (estimate the parameters).
- Find the proportion of these runs in which the estimate is significant.
- This proportion is *empirical power*.

Power analysis in Mplus: Monte Carlo simulation

Say we intend to fit the following random intercept / random slope model.

We want to know how many clusters (and of what size) to aim for to have sufficient power to detect the effect of w_1 on β_{0j} when the rest of the model has been fully specified with reasonable parameter values.

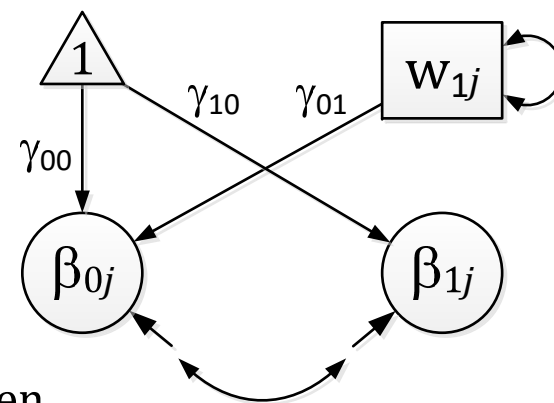


Power analysis in Mplus: Monte Carlo simulation

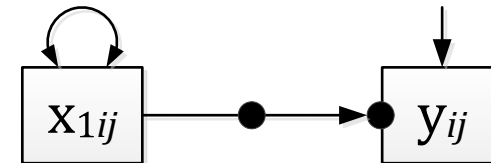
```
TITLE: hsub random intercept and slope (power analysis);
MONTECARLO: NAMES ARE ses mathach sector;
NOOBSERVATIONS = 1000;
CSIZES = 50(20); NREPS = 500;
SEED = 3729; NCSIZES = 1;
WITHIN IS ses; BETWEEN IS sector;
ANALYSIS: TYPE IS TWOLEVEL RANDOM;
MODEL POPULATION:
%WITHIN%
mathach*37; ses*20;
s1 | mathach ON ses;
%BETWEEN%
[mathach*11.5 s1*.4];
mathach*3.9 s1*.5; mathach WITH s1*.7;
mathach ON sector*1.5; sector*.25;
```

MODEL:

```
%WITHIN%
mathach*37; ses*20;
s1 | mathach ON ses;
%BETWEEN%
[mathach*11.5 s1*.4];
mathach*3.9 s1*.5; mathach WITH s1*.7;
mathach ON sector*1.5; sector*.25;
```



Within



Power analysis in Mplus: Monte Carlo simulation

| | Population | ESTIMATES Average | Std. Dev. | S. E. Average | M. S. E. | 95% Cover | % Sig Coeff |
|---------------------|------------|----------------------|-----------|------------------|----------|--------------|----------------|
| Within Level | | | | | | | |
| Means | | | | | | | |
| SES | 0.000 | -0.0023 | 0.1396 | 0.1401 | 0.0195 | 0.934 | 0.066 |
| Variiances | | | | | | | |
| SES | 20.000 | 19.9692 | 0.8801 | 0.8787 | 0.7740 | 0.930 | 1.000 |
| Residual Variiances | | | | | | | |
| MATHACH | 37.000 | 36.9516 | 1.8568 | 1.7077 | 3.4431 | 0.934 | 1.000 |
| Between Level | | | | | | | |
| MATHACH ON | | | | | | | |
| SECTOR | 1.500 | 1.4762 | 0.6717 | 0.6148 | 0.4508 | 0.928 | 0.656 |
| MATHACH WITH | | | | | | | |
| S1 | 0.700 | 0.6768 | 0.2808 | 0.2704 | 0.0792 | 0.932 | 0.750 |

...

Power analysis in Mplus: Monte Carlo simulation

To triangulate on appropriate level-1 and level-2 sample sizes, it may be useful to create a chart like this one:

| J | n_J | power |
|----------|----------------------|--------------|
| 50 | 10 | .52 |
| 75 | 10 | .68 |
| 100 | 10 | .79 |
| 50 | 20 | .66 |
| 75 | 20 | .80 |
| 100 | 20 | .90 |

These two have the same total sample size, but different empirical power estimates.

9. Three-level models

Three-level data

To this point we have focused exclusively on 2-level data.

2-level data can be seen as a special case of 3-level data with only one level-3 unit, or only one level-1 unit within each level-2 unit.

True 3-level data are common. For example:

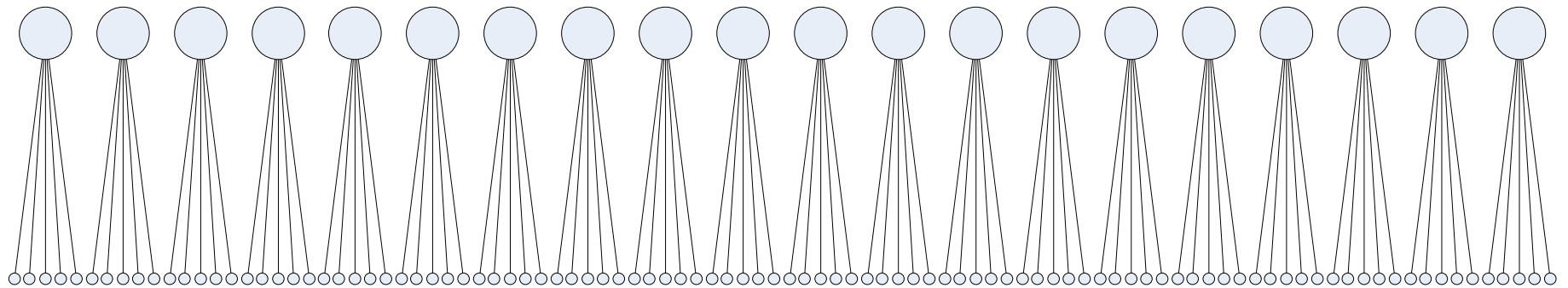
- **employees** within **units** within **organizations**
- **farms** within **counties** within **states**
- **patients** within **physicians** within **hospitals**
- **repeated measures** within **students** within **classrooms**

In addition, multivariate 2-level data are modeled as 3-level data, with an extra “dummy” level for the multivariate outcomes.

As always, the outcome variable is measured at level-1.

Three-level data

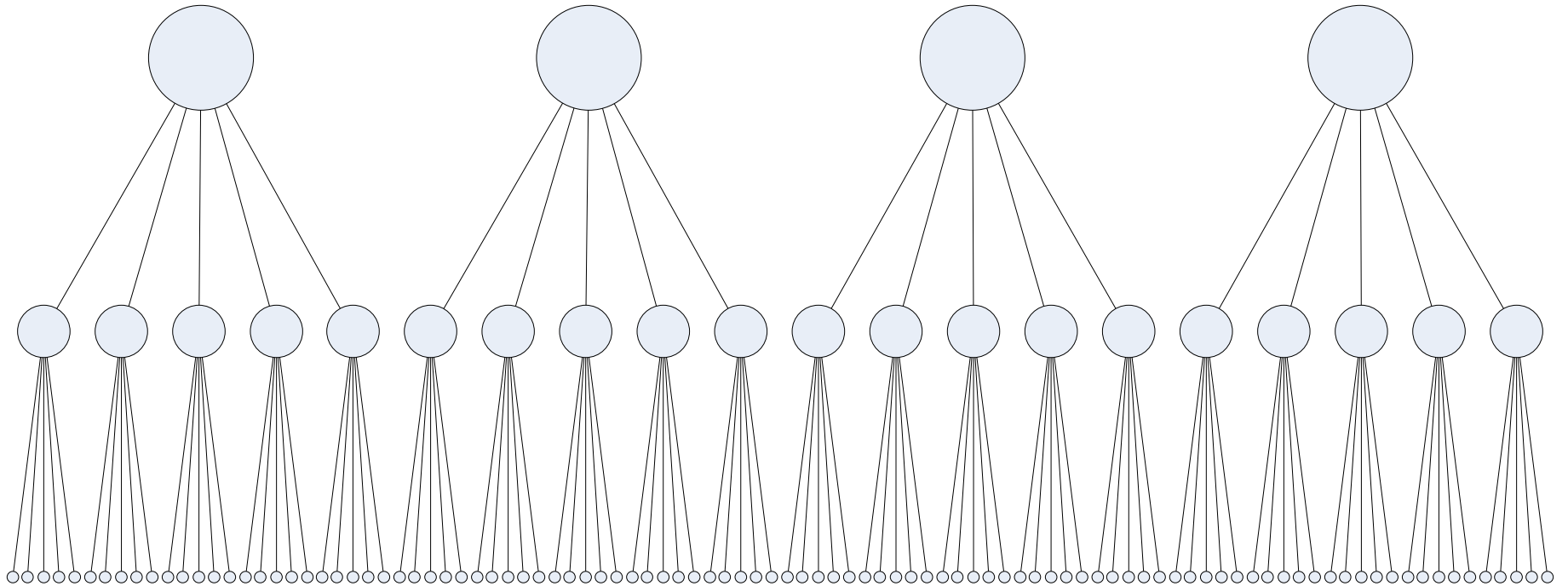
Say you start with a 2-level data structure...



...and you realize that your level-2 units are clustered. Continuing with a 2-level MLM **violates the assumption of independence** for the level-2 u_{ij} residuals.

A third level of nesting must be used.

Three-level data



Adding a third level to model dependence is exactly analogous to using MLM instead of single-level regression to model dependence among residuals.

Three-level models

This kind of model can become complicated **very fast**.

- There are many parameters
- It requires a very large sample
- There is an increased chance of encountering estimation errors

Intercepts and slopes can vary across level-2 units, level-3 units, or both.

Predictors can be added at any level.

Level-1 random intercepts and slopes can be predicted by level-2 and/or level-3 predictors.

Level-2 random intercepts and slopes can be predicted by level-3 predictors.

Interaction effects can exist within any level (1, 2, or 3) or can “cross” any pair of levels.

Three-level models

Level-1, 2, and 3 sample size requirements are very complicated. This represents a largely unexplored area.

It is still true that the number of upper-level units needs to be “large;” that condition may be more difficult to satisfy in 3-level models, depending on the situation.

Some rules of thumb still hold:

- (1) If you are particularly interested in the effect of a level- k predictor, collect a lot of k -level units.
- (2) Upper-level units are typically more important than lower-level units.
- (3) There is no substitute for a good *a priori* power analysis.

Two-level random intercept model (review)

The 2-level null model (random-effects ANOVA):

$$y_{ij} = \beta_{0j} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Reduced-form equation:

$$y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

Residual distributions: $u_{0j} \sim N(0, \tau_{00})$

$$e_{ij} \sim N(0, \sigma_e^2)$$

Three-level random intercept model

The 3-level null model (random-effects ANOVA... sort of):

$$y_{ijk} = \beta_{0jk} + e_{ijk}$$

$$\beta_{0jk} = \beta_{00k} + u_{0jk}$$

$$\beta_{00k} = \gamma_{000} + u_{00k}$$

Reduced-form equation:

$$y_{ijk} = \gamma_{000} + u_{00k} + u_{0jk} + e_{ijk}$$

Residual distributions: $u_{00k} \sim N\left(0, \tau_{00}^{(3)}\right)$

$$u_{0jk} \sim N\left(0, \tau_{00}^{(2)}\right)$$

$$e_{ijk} \sim N\left(0, \sigma_e^2\right)$$

ICC in three-level models

In 2-level null models, the reduced-form equation is:

$$y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

Because these components are independent, the variance of y_{ij} is partitioned as (the variance of the sum is the sum of the variances):

$$\begin{aligned}\hat{\sigma}_y^2 &= \text{var}[\gamma_{00}] + \text{var}[u_{0j}] + \text{var}[e_{ij}] \\ &= 0 + \hat{\tau}_{00} + \hat{\sigma}_e^2 \\ &= \hat{\tau}_{00} + \hat{\sigma}_e^2\end{aligned}$$

The intraclass correlation is therefore:

$$\text{ICC} = \frac{\hat{\tau}_{00}}{\hat{\sigma}_y^2} = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}_e^2}$$

ICC in three-level models

In 3-level null models, the reduced-form equation is:

$$y_{ijk} = \gamma_{000} + u_{00k} + u_{0jk} + e_{ijk}$$

Because these components are independent, the variance of y_{ij} is partitioned as (the variance of the sum is the sum of the variances):

$$\begin{aligned}\hat{\sigma}_y^2 &= \text{var}[\gamma_{000}] + \text{var}[u_{00k}] + \text{var}[u_{0jk}] + \text{var}[e_{ijk}] \\ &= 0 + \hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2 \\ &= \hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2\end{aligned}$$

The intraclass correlation is more complicated in 3-level models.

ICC in three-level models

Recall that in 2-level models, the ICC could be interpreted in two ways:

- The proportion of observed variance that is between units.
- The expected correlation of randomly selected individuals within a group.

These interpretations coincide only when there are 2 levels.

ICC in three-level models

Two methods for computing ICC in 3-level models:

$$ICC_2 = \frac{\hat{\tau}_{00}^{(2)}}{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2}$$

$$ICC_3 = \frac{\hat{\tau}_{00}^{(3)}}{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2}$$

Decomposition of variance.

Identifies the proportion of variance at level-2 and level-3 (and level-1, if we are interested).

$$ICC_2 = \frac{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)}}{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2}$$

$$ICC_3 = \frac{\hat{\tau}_{00}^{(3)}}{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2}$$

Expected correlation between level-1 units in the same level-2 unit.

Expected correlation between level-2 units in the same level-3 unit.

ICC in three-level models

Say you find the following:

$$\frac{\hat{\sigma}_e^2}{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2} = .36$$

$$\frac{\hat{\tau}_{00}^{(2)}}{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2} = .29$$

$$\frac{\hat{\tau}_{00}^{(3)}}{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2} = .35$$

These add up to 1.0 and indicate a nontrivial amount of variance at each level.

ICC in three-level models

These (probably) do not show enough variability to justify using a full 3-level model:

$$\frac{\hat{\sigma}_e^2}{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2} = .02$$

$$\frac{\hat{\sigma}_e^2}{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2} = .24$$

$$\frac{\hat{\tau}_{00}^{(2)}}{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2} = .59$$

$$\frac{\hat{\tau}_{00}^{(2)}}{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2} = .03$$

$$\frac{\hat{\tau}_{00}^{(3)}}{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2} = .39$$

$$\frac{\hat{\tau}_{00}^{(3)}}{\hat{\tau}_{00}^{(3)} + \hat{\tau}_{00}^{(2)} + \hat{\sigma}_e^2} = .73$$

Predictors and random slopes

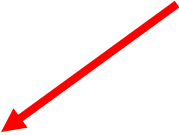
Predictors can be added at any level. For example, level-3:


$$y_{ijk} = \beta_{0jk} + e_{ijk}$$


$$\beta_{0jk} = \beta_{00k} + u_{0jk}$$

$$\beta_{00k} = \gamma_{000} + \gamma_{001}w_k + u_{00k}$$

$$y_{ijk} = \underbrace{\gamma_{000} + \gamma_{001}w_k}_{\text{fixed}} + \underbrace{u_{00k}}_{\text{random3}} + \underbrace{u_{0jk}}_{\text{random2}} + \underbrace{e_{ijk}}_{\text{random1}}$$


$$u_{00k} \sim N(0, \tau_{00}^{(3)})$$


$$u_{0jk} \sim N(0, \tau_{00}^{(2)})$$


$$e_{ijk} \sim N(0, \sigma_e^2)$$

Predictors and random slopes

Predictor added at level-2 (slopes random at level-3):

$$y_{ijk} = \beta_{0jk} + e_{ijk}$$

$$\beta_{0jk} = \beta_{00k} + \beta_{01k} w_{jk} + u_{0jk}$$

$$\beta_{00k} = \gamma_{000} + u_{00k}$$

$$\beta_{01k} = \gamma_{010} + u_{01k}$$

$$u_{0jk} \sim N(0, \tau_{00}^{(2)})$$

$$e_{ijk} \sim N(0, \sigma_e^2)$$

$$y_{ijk} = \underbrace{\gamma_{000} + \gamma_{010} w_{jk}}_{\text{fixed}} + \underbrace{u_{00k} + u_{01k} w_{jk}}_{\text{random3}} + \underbrace{u_{0jk}}_{\text{random2}} + \underbrace{e_{ijk}}_{\text{random1}}$$

$$\begin{bmatrix} u_{00k} \\ u_{01k} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00}^{(3)} & \\ & \tau_{11}^{(3)} \end{bmatrix} \right)$$

Predictors and random slopes

Predictor added at level-1 (slopes random at level-2 *and* level-3):

$$y_{ijk} = \beta_{0jk} + \beta_{1jk} x_{ijk} + e_{ijk}$$

$$\beta_{0jk} = \beta_{00k} + u_{0jk}$$

$$\beta_{1jk} = \beta_{10k} + u_{1jk}$$

$$\beta_{00k} = \gamma_{000} + u_{00k}$$

$$\beta_{10k} = \gamma_{100} + u_{10k}$$

$$\begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00}^{(2)} & \\ & \tau_{11}^{(2)} \end{bmatrix} \right)$$

$$e_{ijk} \sim N(0, \sigma_e^2)$$

$$y_{ijk} = \underbrace{\gamma_{000} + \gamma_{100} x_{ijk}}_{\text{fixed}} + \underbrace{u_{00k} + u_{10k} x_{ijk}}_{\text{random3}} + \underbrace{u_{0jk} + u_{1jk} x_{ijk}}_{\text{random2}} + \underbrace{e_{ijk}}_{\text{random1}}$$

$$\begin{bmatrix} u_{00k} \\ u_{10k} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00}^{(3)} & \\ & \tau_{11}^{(3)} \end{bmatrix} \right)$$

Interactions in three-level models

Say we have a 3-level model with one predictor, random intercepts and slopes at both level-2 and level-3, and a level-3 predictor of those random intercepts and slopes:

$$y_{ijk} = \beta_{0jk} + \beta_{1jk}x_{ijk} + e_{ijk}$$

$$\beta_{0jk} = \beta_{00k} + u_{0jk}$$

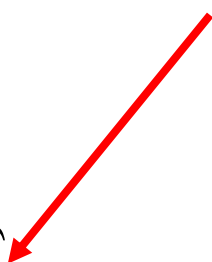
$$\beta_{1jk} = \beta_{10k} + u_{1jk}$$

$$\beta_{00k} = \gamma_{000} + \gamma_{001}w_k + u_{00k}$$

$$\beta_{10k} = \gamma_{100} + \gamma_{101}w_k + u_{10k}$$

$$y_{ijk} = \underbrace{\gamma_{000} + \gamma_{100}x_{ijk} + \gamma_{001}w_k + \gamma_{101}x_{ijk}w_k}_{\text{fixed}} + \underbrace{u_{00k} + u_{0jk} + u_{10k}x_{ijk} + u_{1jk}x_{ijk} + e_{ijk}}_{\text{random}}$$

cross-level
interaction term



Interactions in three-level models

Or, say we add a level-2 predictor to the mix:

$$y_{ijk} = \beta_{0jk} + \beta_{1jk} x_{ijk} + e_{ijk}$$

$$\beta_{0jk} = \beta_{00k} + \beta_{01k} w_{jk} + u_{0jk}$$

$$\beta_{1jk} = \beta_{10k} + \beta_{11k} w_{jk} + u_{1jk}$$

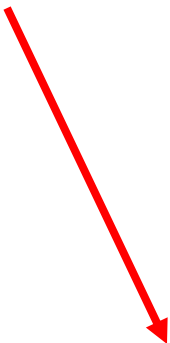
$$\beta_{00k} = \gamma_{000} + \gamma_{001} w_k + u_{00k}$$

$$\beta_{10k} = \gamma_{100} + \gamma_{101} w_k + u_{10k}$$

$$\beta_{01k} = \gamma_{010} + \gamma_{011} w_k + u_{01k}$$

$$\beta_{11k} = \gamma_{110} + \gamma_{111} w_k + u_{11k}$$

three-way
cross-level
interaction term



$$y_{ijk} = \underbrace{\gamma_{000} + \gamma_{100} x_{ijk} + \gamma_{010} w_{jk} + \gamma_{001} w_k + \gamma_{110} x_{ijk} w_{jk} + \gamma_{101} x_{ijk} w_k + \gamma_{011} w_{jk} w_k + \gamma_{111} x_{ijk} w_k w_{jk}}_{\text{fixed}} + \underbrace{u_{00k} + u_{01k} w_{jk} + u_{0jk} + u_{10k} x_{ijk} + u_{11k} x_{ijk} w_{jk} + u_{1jk} x_{ijk}}_{\text{random}} + e_{ijk}$$

Interactions in three-level models

$$y_{ijk} = \underbrace{\gamma_{000} + \gamma_{100}x_{ijk} + \gamma_{001}w_k + \gamma_{101}x_{ijk}w_k}_{\text{fixed}} + \underbrace{u_{00k} + u_{0jk} + u_{10k}x_{ijk} + u_{1jk}x_{ijk} + e_{ijk}}_{\text{random}}$$

$$y_{ijk} = \underbrace{\gamma_{000} + \gamma_{100}x_{ijk} + \gamma_{010}w_{jk} + \gamma_{001}w_k + \gamma_{110}x_{ijk}w_{jk} + \gamma_{101}x_{ijk}w_k + \gamma_{011}w_{jk}w_k + \gamma_{111}x_{ijk}w_kw_{jk}}_{\text{fixed}} + \underbrace{u_{00k} + u_{01k}w_{jk} + u_{0jk} + u_{10k}x_{ijk} + u_{11k}x_{ijk}w_{jk} + u_{1jk}x_{ijk} + e_{ijk}}_{\text{random}}$$

Even though 3-level models with interaction terms may *appear* complicated, remember that only the fixed part of the model is used to plot and probe interaction effects, and it appears **exactly the same as in the 2-level model**.

The only difference is in the random effects, which are irrelevant for interactions.

The same web page calculators can be used regardless of how many levels there are.

Consequences of ignoring a level

Moerbeek (2004) explored what would happen if a level (either level-2 or level-3) were ignored when the data are really structured in 3 levels.

The variance at the missing level is “redistributed” in predictable ways.

When level-3 is ignored (e.g., when students are modeled as nested within classes, and school is ignored):

- The level-1 residual variance does not change.
- The real level-3 variance is added to the level-2 variance.

$$\hat{\tau}_{00}^{(2)} = \hat{\tau}_{00}^{(2)} + \hat{\tau}_{00}^{(3)}$$

$$\hat{\sigma}_e^2 = \hat{\sigma}_e^2$$

- Level-2 slope estimates have inflated *SEs* (and thus lower power), but there is no effect on level-1 slopes.

Consequences of ignoring a level

When level-2 is ignored (e.g., when students are modeled as nested within schools, and class is ignored):

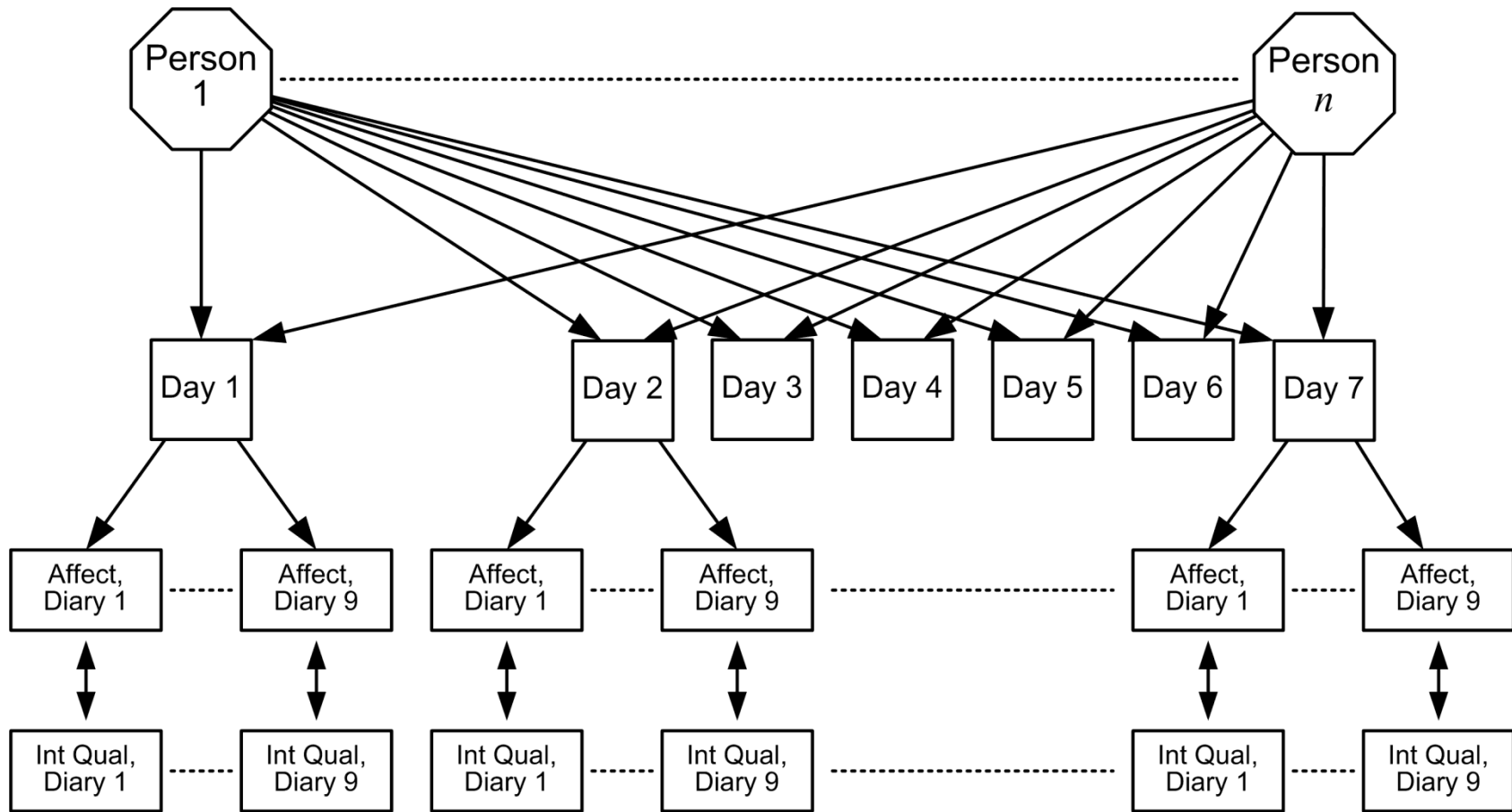
- The level-2 variance is redistributed to both level-1 and level-3 (now level-2).

$$\hat{\tau}_{00}^{(3)} = \hat{\tau}_{00}^{(2)} + \left(\frac{n_1 - 1}{n_1 n_2 - 1} \right) \hat{\tau}_{00}^{(1)}$$

$$\hat{\sigma}_e^2 = \hat{\sigma}_e^2 + \left(\frac{n_1 n_2 - n_1}{n_1 n_2 - 1} \right) \hat{\tau}_{00}^{(1)}$$

- Level-1 slope estimates have inflated *SEs*.

Hawkey et al. (2007) Loneliness data



Models for Hawkley et al.'s (2007) loneliness data

The supplementary materials contain code in R, Mplus, and SPSS for 7 models using positive affect (POSAFF) as the outcome:

1. Null model
2. Lag-1 interaction quality, fixed slope
3. Lag-1 interaction quality, random slope at levels 1 and 3
4. Weekend, fixed slope
5. Weekend, random slope at level 3
6. Loneliness, fixed slope
7. Loneliness and weekend interaction, random slope for weekend at level 3

Beyond three levels

Of course, more than three levels are possible. MLwiN can handle up to 100 levels (!).

Modeling more than three levels taxes not only software capacity, but readers' ability to understand and interpret parameters.

Using more than three levels requires strong theoretical support.

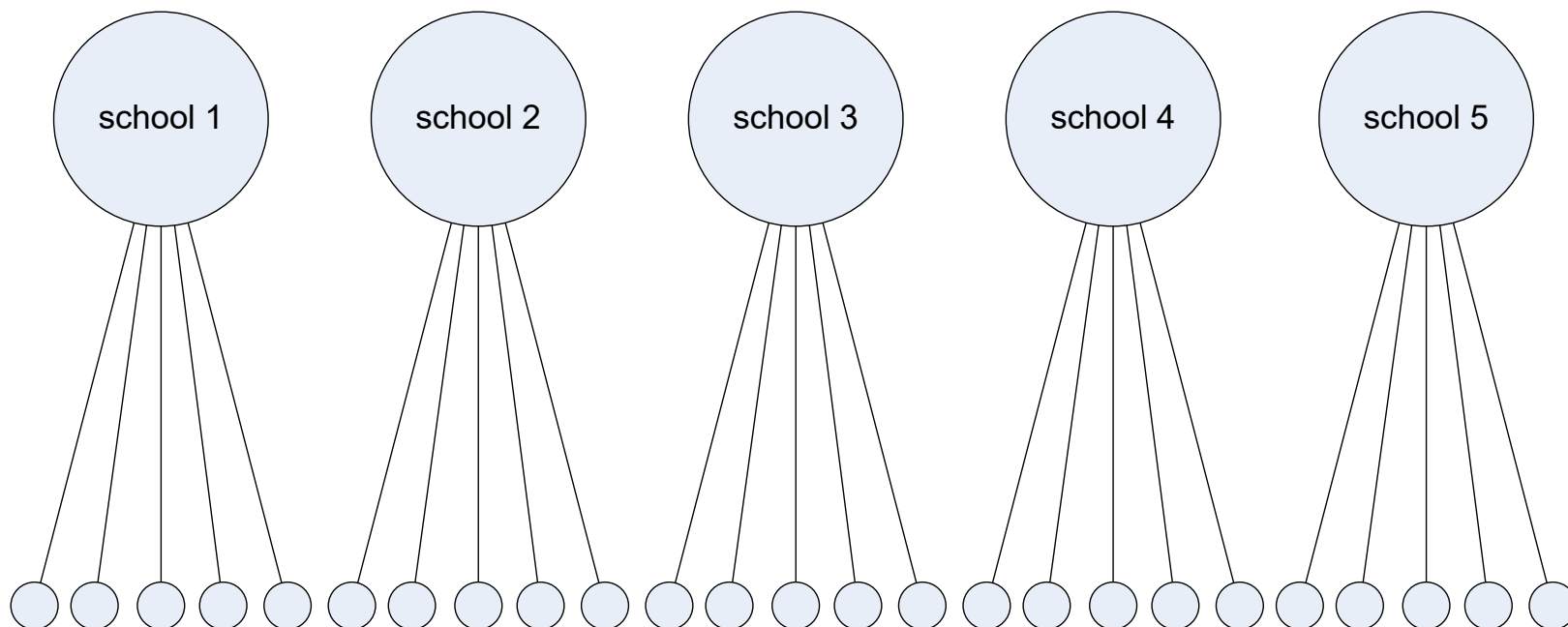
10. A model for cross-classified data

Nonhierarchical data

To date, the data structures we have examined have been cleanly hierarchical. For example:

Students nested within schools. A given student is in only one school. Data from different schools are collected from different sources and are independent.

Repeated measures within persons. Repeated measurements, by their very nature, are nested within one person. Different people have different repeated measurements.



Nonhierarchical data

Data are not always so cleanly nested. A common way in which data can be “quasi-hierarchical” (vs. nonhierarchical) is **cross-classification**: What if subjects are nested in two *kinds* of level-2 units?

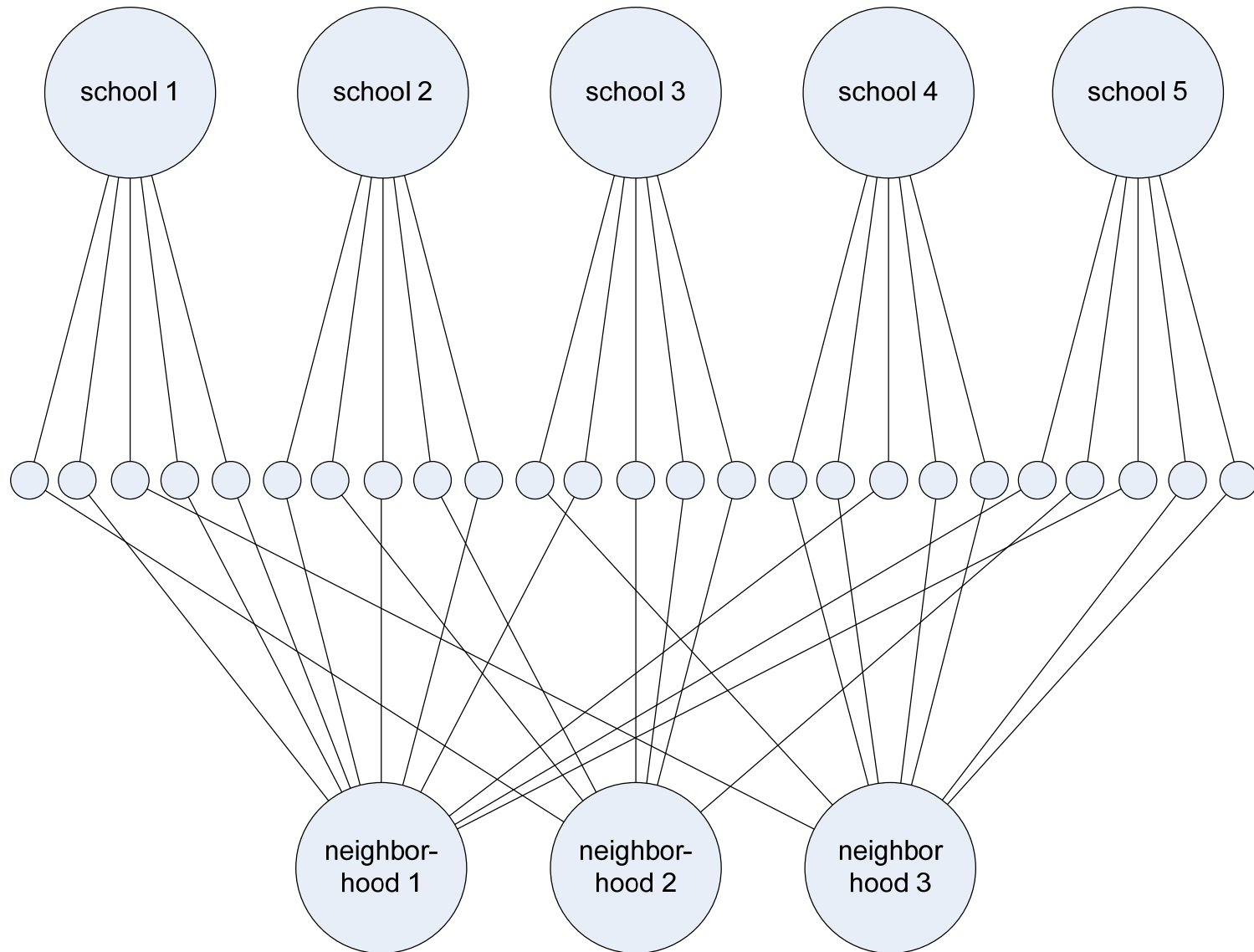
In most cases of cross-classification (CC), there is a 2-way classification at Level-2.

The classic example of CC: students nested within **both schools and neighborhoods**.

Students may be grouped in terms of the schools they attend **or** the neighborhoods where they live.

A given pair of students may attend the same school but live in different neighborhoods, or vice versa. Both are relevant clusters when considering lack of independence of scores.

Visualizing cross-classification



Visualizing cross-classification

One way to visualize a cross-classification:

| School | Neighborhood | | | | | | | |
|--------|--------------|----|---|----|----|----|----|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 12 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 2 | 3 | 13 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 4 | 0 | 9 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 2 | 4 | 14 | 0 | 1 | 0 | 0 |
| 5 | 0 | 0 | 0 | 1 | 15 | 5 | 2 | 2 |
| 6 | 0 | 0 | 1 | 0 | 2 | 16 | 1 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 1 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |

Visualizing cross-classification

Completely confounded classifications (no cross-classified model possible):

| Class | Teacher | | | | | | | |
|-------|---------|----|----|---|---|----|----|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 15 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 6 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 11 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 11 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |

Visualizing cross-classification

Classes nested within schools:

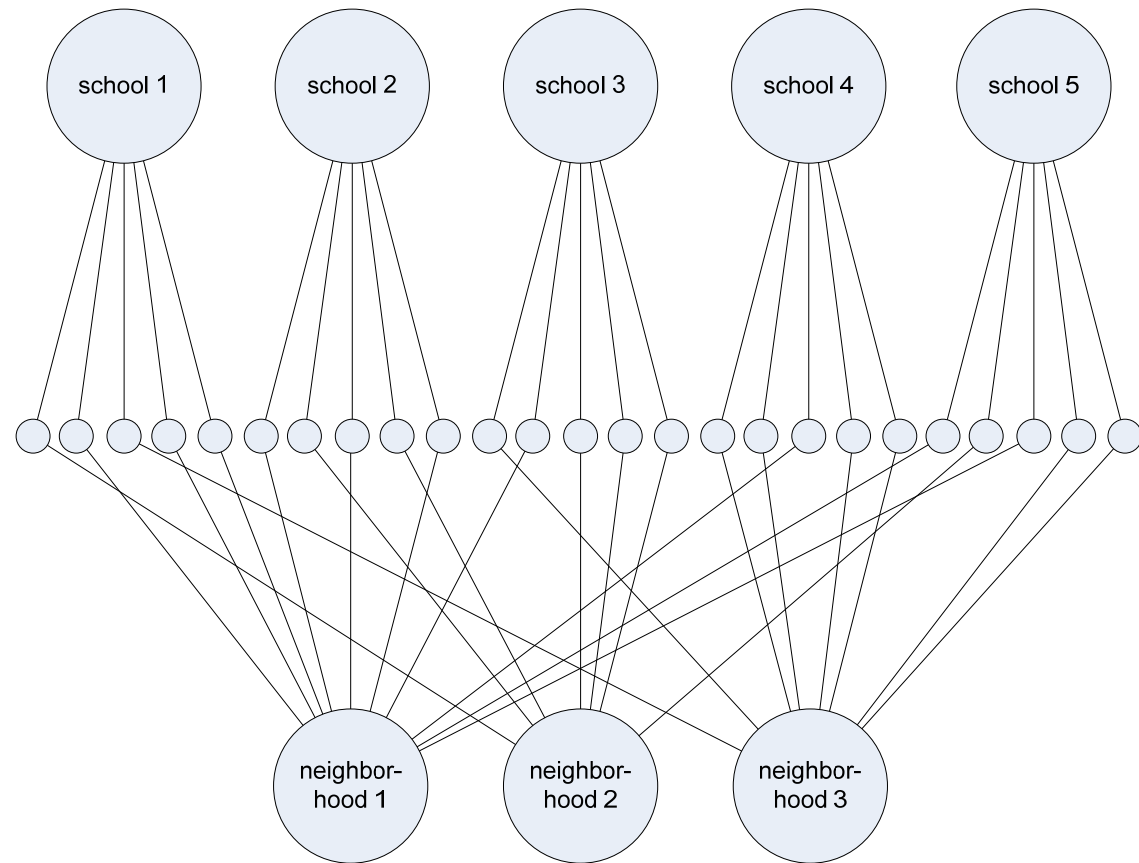
| School | Class | | | | | | | |
|--------|-------|----|----|----|----|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 14 | 17 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 12 | 14 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 14 | 8 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 4 |

Cross-classified data

Students within schools are more similar to each other than to students in other schools. We should therefore specify a random intercept for schools.

Students within neighborhoods are more similar to each other than to students in other neighborhoods. We should therefore specify a random intercept for neighborhoods.

Students within the same neighborhood *and* school are, in theory, *especially* similar. We should therefore specify a random intercept for school \times neighborhood.



One-way random effects ANOVA

Taking a step back...

One-way random-effects ANOVA (RANOVA) model:

The model:

$$y_{ij} = \beta_{0j} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

We assume:

$$e_{ij} \sim N(0, \sigma_e^2)$$

$$u_{0j} \sim N(0, \tau_{00})$$

$$\text{cov}(u_{0j}, e_{ij}) = 0$$

One-way random effects ANOVA

In one-way RANOVA models, the reduced-form equation is:

$$y_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

Because these components are independent, the variance of y_{ij} is partitioned as (the variance of the sum is the sum of the variances):

$$\begin{aligned}\hat{\sigma}_y^2 &= \text{var}[\gamma_{00}] + \text{var}[u_{0j}] + \text{var}[e_{ij}] \\ &= 0 + \hat{\tau}_{00} + \hat{\sigma}_e^2 \\ &= \hat{\tau}_{00} + \hat{\sigma}_e^2\end{aligned}$$

The intraclass correlation is therefore:

$$\text{ICC} = \frac{\hat{\tau}_{00}}{\hat{\sigma}_y^2} = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}_e^2}$$

Two-way (factorial) random effects ANOVA

Two-way random-effects ANOVA (RANOVA) model:

The model:

$$y_{i(jk)} = \beta_{0(jk)} + e_{i(jk)}$$

$$\beta_{0(jk)} = \gamma_{00} + \underbrace{u_{0j}}_{\text{school}} + \underbrace{u_{0k}}_{\text{neighborhood}} + \underbrace{u_{0(jk)}}_{\text{school} \times \text{neighborhood}}$$

$$y_{i(jk)} = \gamma_{00} + \underbrace{u_{0j} + u_{0k} + u_{0(jk)}}_{\text{crossed random effects at Level-2}} + e_{i(jk)}$$

We assume:

$$e_{i(jk)} \sim N(0, \sigma_e^2)$$

$$u_{0j} \sim N(0, \tau_{00}^{(j)})$$

$$u_{0k} \sim N(0, \tau_{00}^{(k)})$$

$$u_{0(jk)} \sim N(0, \tau_{00}^{(jk)})$$

all residuals
are independent

This term usually is omitted for practical reasons, so I will omit it here.

Estimation considerations

Usually the jk component is omitted for parsimony or due to lack of data.

Sometimes the cross-classification will be too sparse to estimate a jk random effect. In such cases we can still estimate a “main effects only” model.

| School | Neighborhood | | | | | | | |
|--------|--------------|----|---|----|----|----|----|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 12 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 2 | 3 | 13 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 4 | 0 | 9 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 2 | 4 | 14 | 0 | 1 | 0 | 0 |
| 5 | 0 | 0 | 0 | 1 | 15 | 5 | 2 | 2 |
| 6 | 0 | 0 | 1 | 0 | 2 | 16 | 1 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 1 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |

Two-way (factorial) random effects ANOVA

In two-way RANOVA models, the reduced-form equation is:

$$y_{i(jk)} = \gamma_{00} + u_{0j} + u_{0k} + e_{i(jk)}$$

Because these components are independent, the variance of y_{ij} is partitioned as (the variance of the sum is the sum of the variances):

$$\begin{aligned}\hat{\sigma}_y^2 &= \text{var}[\gamma_{00}] + \text{var}[u_{0j}] + \text{var}[u_{0k}] + \text{var}[e_{i(jk)}] \\ &= 0 + \hat{\tau}_{00}^{(j)} + \hat{\tau}_{00}^{(k)} + \hat{\sigma}_e^2 \\ &= \hat{\tau}_{00}^{(j)} + \hat{\tau}_{00}^{(k)} + \hat{\sigma}_e^2\end{aligned}$$

There is no longer only one ICC.

ICC in two-way RANOVA models

In terms of proportion of variance (these correspond roughly to ANOVA main effects and interactions)...

$$\text{ICC}_j = \frac{\hat{\tau}_{00}^{(j)}}{\hat{\tau}_{00}^{(j)} + \hat{\tau}_{00}^{(k)} + \hat{\sigma}_e^2}$$

is the proportion of total observed variance that is due uniquely to sharing the same school (j).

$$\text{ICC}_k = \frac{\hat{\tau}_{00}^{(k)}}{\hat{\tau}_{00}^{(j)} + \hat{\tau}_{00}^{(k)} + \hat{\sigma}_e^2}$$

is the proportion of total observed variance that is due uniquely to sharing the same neighborhood (k).

ICC in two-way RANOVA models

In terms of correlations among level-1 units...

$$ICC_j = \frac{\hat{\tau}_{00}^{(j)}}{\hat{\tau}_{00}^{(j)} + \hat{\tau}_{00}^{(k)} + \hat{\sigma}_e^2}$$

is the correlation among students in the same school (j) but different neighborhoods.

$$ICC_k = \frac{\hat{\tau}_{00}^{(k)}}{\hat{\tau}_{00}^{(j)} + \hat{\tau}_{00}^{(k)} + \hat{\sigma}_e^2}$$

is the correlation among students from the same neighborhood (k) but who attend different schools.

$$ICC_{jk} = \frac{\hat{\tau}_{00}^{(j)} + \hat{\tau}_{00}^{(k)}}{\hat{\tau}_{00}^{(j)} + \hat{\tau}_{00}^{(k)} + \hat{\sigma}_e^2}$$

is the correlation among students who share the same neighborhood and school (jk), or $ICC_j + ICC_k$.

Adding predictors to two-way RANOVA models

We can add predictors at any level. For example, level-2 predictors:

$$y_{i(jk)} = \beta_{0(jk)} + e_{i(jk)}$$

$$\beta_{0(jk)} = \gamma_{00} + \gamma_{01}w_{1j} + \gamma_{02}w_{2k} + u_{0j} + u_{0k}$$

$$y_{i(jk)} = \underbrace{\gamma_{00} + \gamma_{01}w_{1j} + \gamma_{02}w_{2k}}_{\text{fixed portion}} + \underbrace{u_{0j} + u_{0k}}_{\text{crossed random effects at Level-2}} + e_{i(jk)}$$

An odd consequence of having two level-2 classifications (j and k) is that Level-2 predictors measured with respect to j can have level-2 slopes that vary over k , and vice versa. We can even have **cross-classified interaction effects**:

$$y_{i(jk)} = \underbrace{\gamma_{00} + \gamma_{01}w_{1j} + \gamma_{02}w_{2k} + \gamma_{03}w_{1j}w_{2k}}_{\text{fixed portion}} + \underbrace{u_{0j} + u_{0k}}_{\text{crossed random effects at Level-2}} + e_{i(jk)}$$

Adding predictors to two-way RANOVA models

Or we can add level-1 predictors:

$$y_{i(jk)} = \beta_{0(jk)} + \beta_{1(jk)}x_{i(jk)} + e_{i(jk)}$$

$$\beta_{0(jk)} = \gamma_{00} + u_{0j} + u_{0k}$$

$$\beta_{1(jk)} = \gamma_{10} + u_{1j} + u_{1k}$$

$$y_{i(jk)} = \gamma_{00} + \gamma_{10}x_{i(jk)} + \underbrace{u_{0j} + u_{0k}}_{\substack{\text{residuals for school,} \\ \text{neighborhood, and} \\ \text{school} \times \text{neighborhood}}} + \underbrace{u_{1j}x_{i(jk)} + u_{1k}x_{i(jk)}}_{\substack{\text{slope residuals for school, neighborhood,} \\ \text{and school} \times \text{neighborhood}}} + e_{i(jk)}$$

Remember that residuals for j and k are independent. Thus, the \mathbf{T} matrix would be structured:

$$\begin{bmatrix} u_{0j} \\ u_{1j} \\ u_{0k} \\ u_{1k} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00}^{(j)} & & & \\ \tau_{10}^{(j)} & \tau_{11}^{(j)} & & \\ 0 & 0 & \tau_{00}^{(k)} & \\ 0 & 0 & \tau_{10}^{(k)} & \tau_{11}^{(k)} \end{bmatrix} \right)$$

school

neighborhood

How to specify and run two-way RANOVA models

In SAS, a RANOVA model with no predictors would be:

```
proc mixed data=data_file method=reml noclprint;  
  class school neighborhood schlhood  
  model y = / solution alpha=.05 ddfm=kr;  
  random intercept / subject=school;  
  random intercept / subject=neighborhood;  
  random intercept / subject=schlhood;  
run;
```

optional

...and predictors are easy to add. Very simple. PROC MIXED was programmed with ANOVA-like models in mind, not MLM. This gives it the versatility necessary to handle complex variance decompositions.

HLM has a CC feature, too.

MLwiN has no packaged capability to run CC models. It does contain a SETX command to help generate dummy variables for a “trick” procedure to be discussed shortly.

LISREL also does not have a canned capability for running CC models.

How to specify and run two-way RANOVA models

Cross-classified models are also easy to specify using SPSS syntax.

Each nesting unit has its own 'Random' statement.

If the data are not too sparse, the interaction term can also have its own 'Random' statement.

```
MIXED achievement BY school neighborhood  
/FIXED=INTERCEPT  
/METHOD=ML  
/PRINT=SOLUTION TESTCOV  
/RANDOM=INTERCEPT | SUBJECT(school) COVTYPE(VC)  
/RANDOM=INTERCEPT | SUBJECT(neighborhood) COVTYPE(VC)  
/RANDOM=INTERCEPT | SUBJECT(school*neighborhood) COVTYPE(VC).
```

Note: Cross-classified models can take a long time to converge in SPSS if the classification interaction term (school*neighborhood) is included.

How to specify and run two-way RANOVA models

SPSS MIXED, SAS (PROC MIXED), lmer (R), and now Mplus are currently the software of choice for such models.

How to specify and run two-way RANOVA models

The data for this example are from 3,435 children in Scotland who attended 148 primary schools and 19 secondary schools.

The following variables are used for the example:

ATTAIN: Attainment score of pupils at age 16

PID: Primary school identifying code

SEX: Pupil's gender 0 = boy 1 = girl

SID: Secondary school identifying code

Reference:

Paterson, L. (1991). Socio economic status and educational attainment: a multidimensional and multilevel study. *Evaluation and Research in Education*, 5, 97-121.

Note: It took about 1 hour for SPSS to estimate this 'simple' CC model.

Without the interaction term, it takes less than 1 second in SPSS or Mplus.

Consequences of omitting a classification

In general, omitting a classification has the same effects as omitting a level, as discussed earlier.

The model will be misspecified, significance of fixed effects will be spuriously inflated, random effect variances will be all over the map, etc.

In situations where there is meaningful variance in both classifications, *and* the classifications are correlated (they usually are), *and* one classification (say, schools) is omitted, then the variance components for the *other* classification (say, neighborhoods) will tend to be inflated.

Extensions

It is possible to have cross-classifications at Level-1 (e.g., a single observations nested within a cross-classification of stimulus and repeated measure, both nested in children). See Hox' (2002) second example in Chapter 7 using network peer-rating data.

MMMC (“multiple membership multiple classification”) models combine cross-classification with multiple membership.

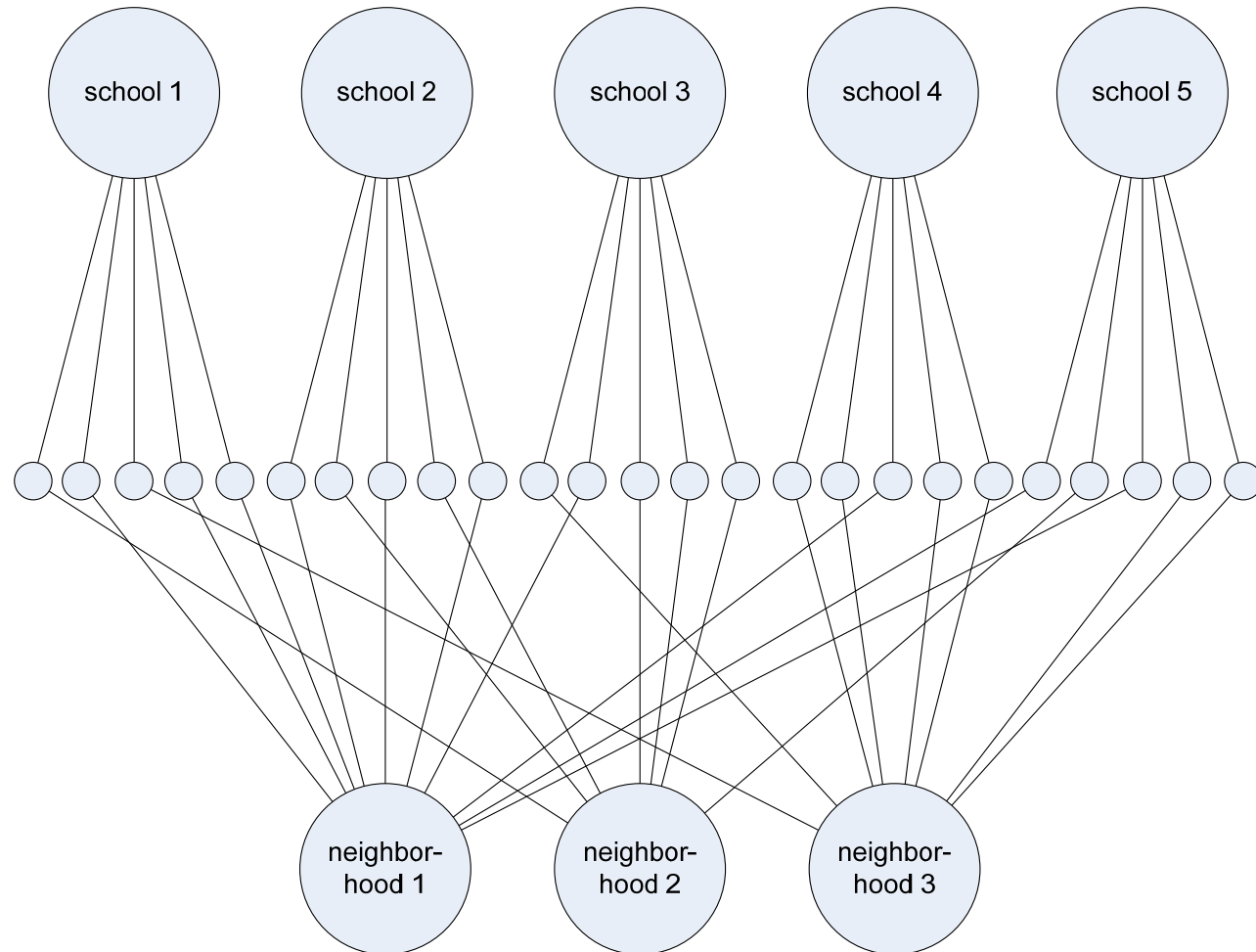
Three-way (or more) cross-classifications are possible (e.g., school × neighborhood × rater).

Partial three-way cross-classifications are possible (e.g., school × neighborhood, where rater is crossed with neighborhood but not with school).

Classification diagrams

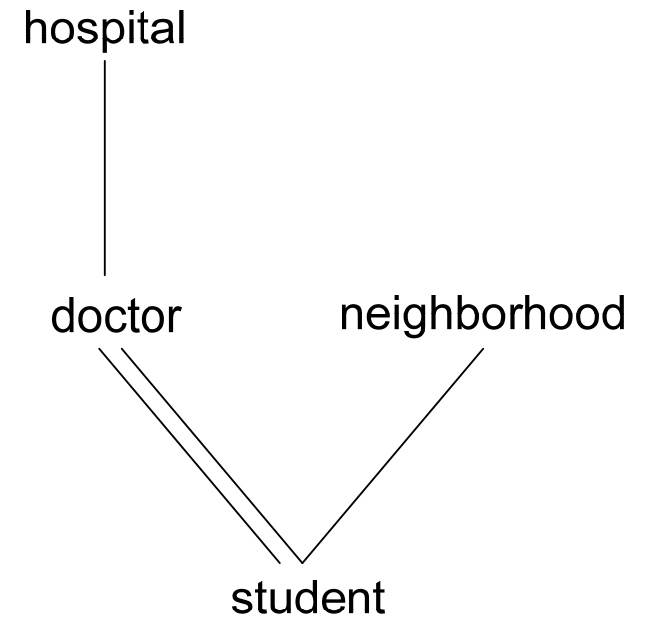
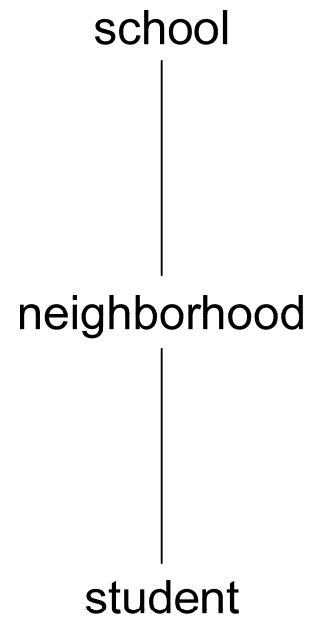
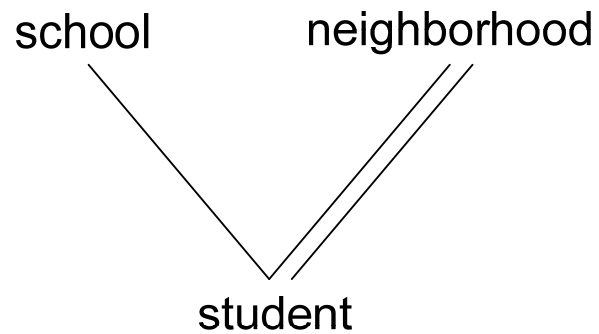
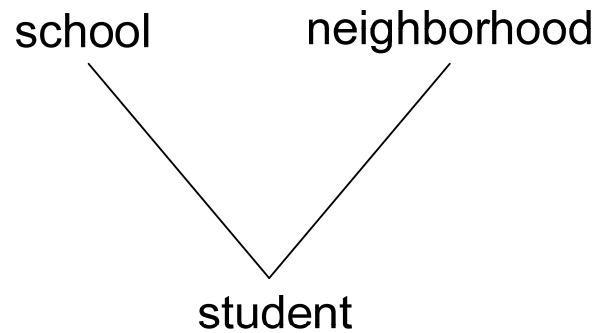
It can be difficult to understand or convey the nesting structure of complex data sets. Rasbash and Browne recommend using **classification diagrams**.

Rather than...



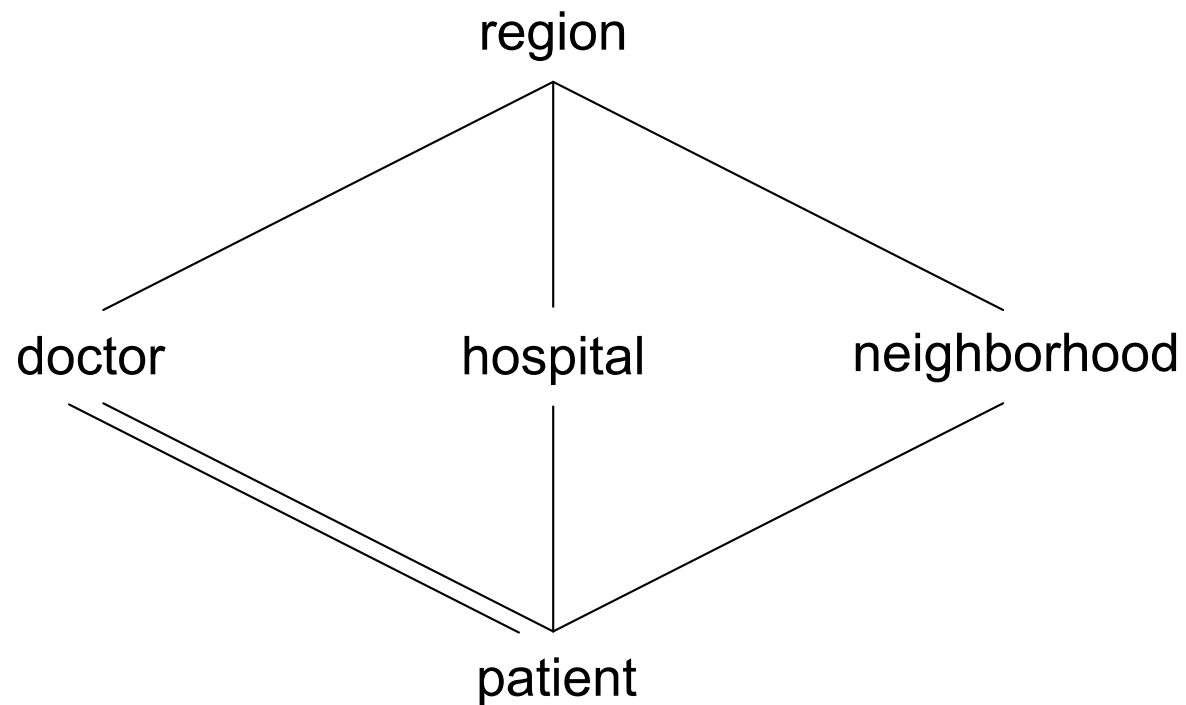
Classification diagrams

Consider using classification diagrams. These are simple and unambiguous:



Classification diagrams

...and can be used to represent very complex sampling schemes:



Complex classification

Some of these designs can be fit using ordinary multilevel software. Most cannot.

For data with complex classification, consider using Paras Mehta's R program **xxM**.

11. Models for categorical outcomes

Categorical, count, censored, and other kinds of outcomes

Up to this point, we have been assuming either univariate or multivariate dependent variables with continuous, normal distributions.

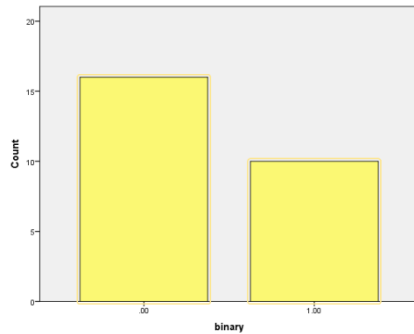
The multilevel linear model is generally appropriate when the outcome is continuous and normal.

Normality is approximately satisfied when the outcome is (for instance) a summative scale score, because of the Central Limit Theorem.

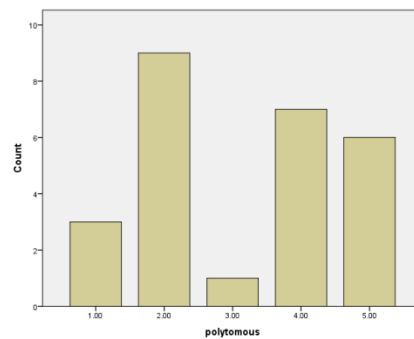
But there are many other kinds of data out there that may not even approximately satisfy the assumption of normality...

Categorical, count, censored, and other kinds of outcomes

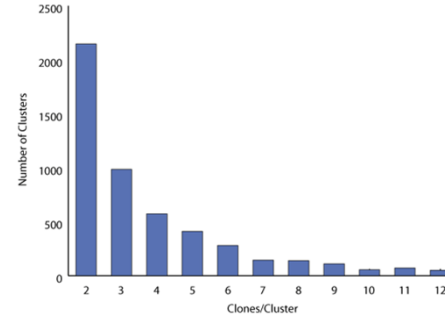
Examples of outcome distributions that are...



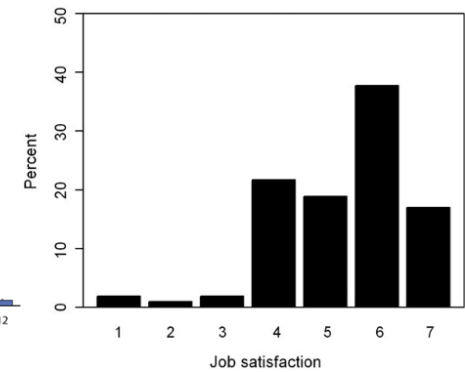
Binary / dichotomous



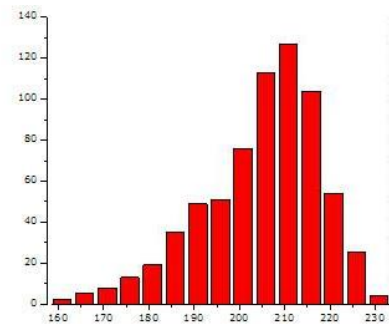
Categorical /
Polytomous



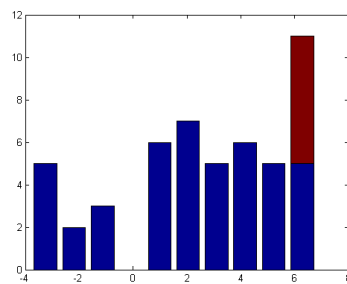
Count



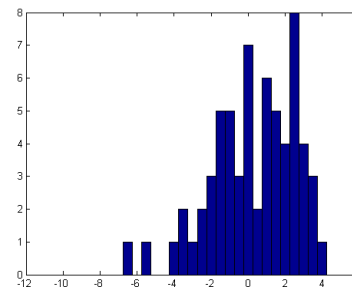
Ordered categorical



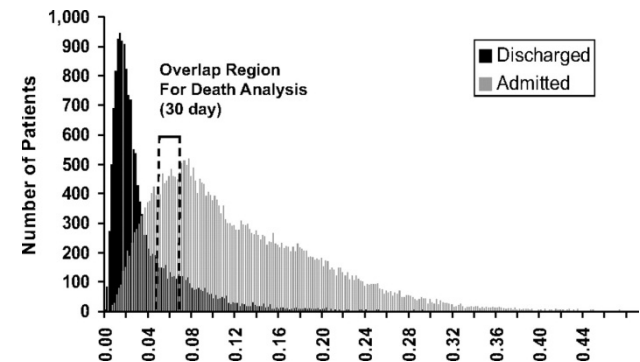
Skewed



Censored



Truncated



Time-to-event

Categorical, count, censored, and other kinds of outcomes

When the outcome violates normality in some way, or is discrete rather than continuous, we cannot use the standard linear model.

Why not?

Why linear models are not always appropriate

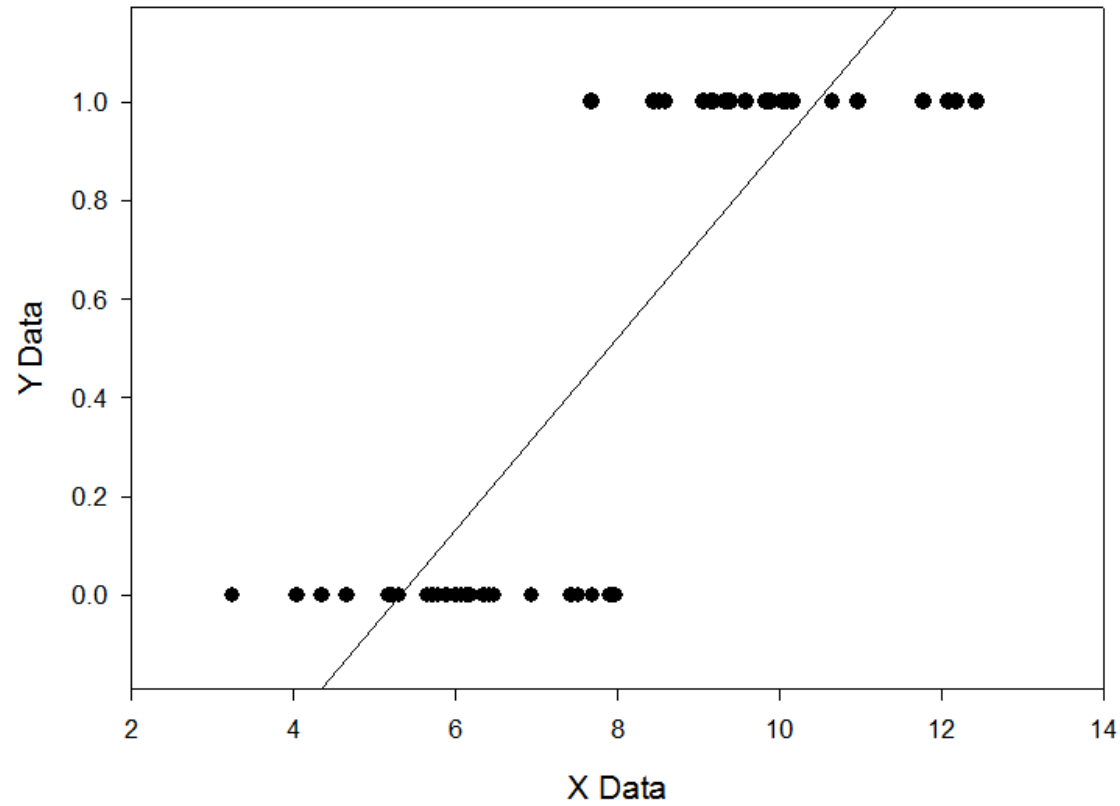
The problem is that such outcomes violate some distributional assumptions that are necessary for proper estimation.

Consider binary outcomes $(0,1)$.

- Normality of the errors (and therefore of y) is violated.
- Homoscedasticity of the errors is violated.
- The model implies nonsensical predicted values of y (violation of linearity).

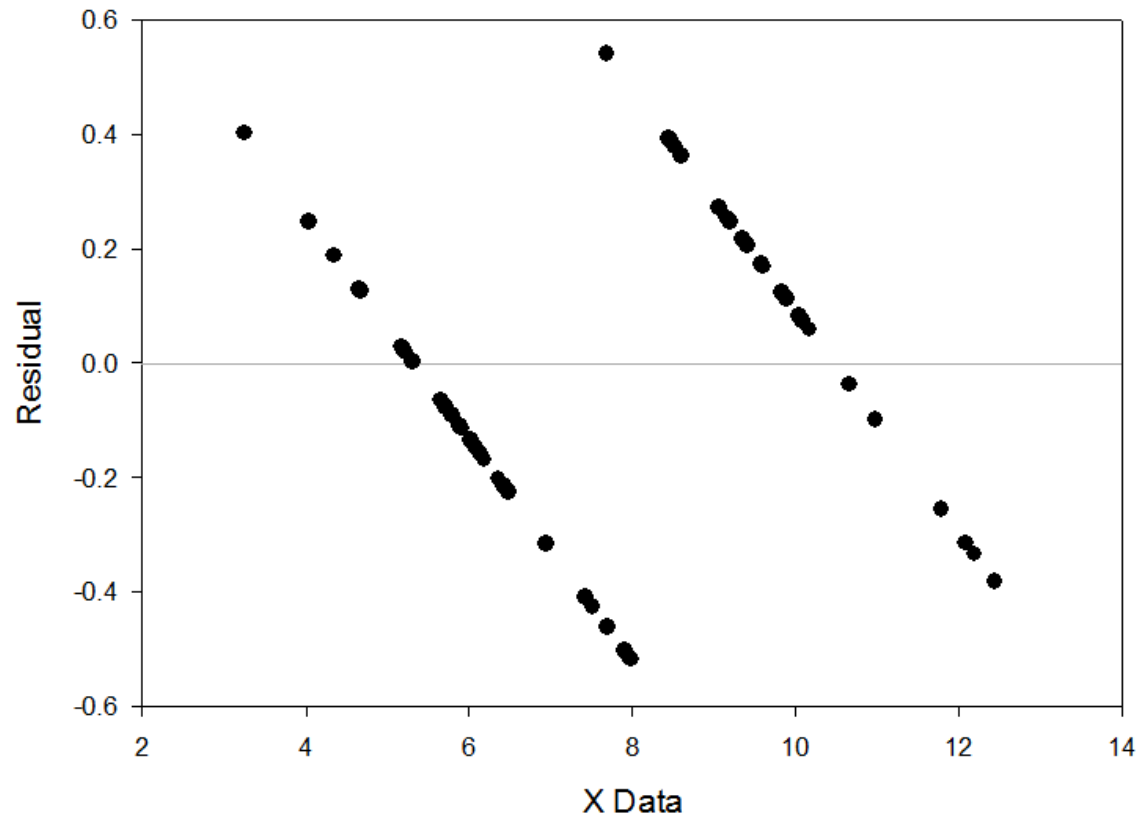
Why linear models are not always appropriate

Distribution of binary y plotted against x ...



Why linear models are not always appropriate

Distribution of residuals plotted against x ...



Similar issues occur with other outcome types.

Data transformation

What can we do in such situations?

One common strategy is to transform x , y , or both in an attempt to bring them into line with assumptions.

Skewed or count data can often be transformed using a square-root (\sqrt{x}), logarithmic ($\ln(x)$), or negative reciprocal ($-1/x$) transformation, then applying usual regression methods.

Transformation is problematic:

- Changes the response scale such that it is no longer interval-scaled (often).
- The results may be uninterpretable, or not as interpretable as before transformation.
- They are ad hoc.
- Binary responses cannot be transformed into anything resembling normality.

I would avoid transformation if possible.

Modeling discrete responses

The modern solution to these problems is to acknowledge the nonnormal nature of such DVs, and to explicitly model them in a way that takes nonnormality into account by choosing an appropriate error distribution.

This requires a class of models known as **generalized linear models** (GLM).

Generalized linear modeling (GLM)

The **generalized linear model (GLM)** has three components:

- The **response distribution**.^{*} The outcome variable y is assumed to have a specific distribution that usually has a mean (μ) and an error variance (σ^2). For example, normal, Bernoulli, negative binomial, gamma, beta, Poisson...
- The **linear predictor**. A linear additive equation that produces an unobserved predictor (η) of the outcome y . Optimal linear combination of covariates (a regression equation).
- The **link function**. A function that links the expected value of the outcome y (i.e., μ) to the predicted values for η . Must be monotonic. For example, logit, probit, log...

^{*}also called the *error distribution* or *probability distribution*.

Generalized linear modeling (GLM)

Ordinary **multiple linear regression** is a special case of GLM:

- The **response distribution** of y is assumed normal with mean μ and variance σ^2 .

$$y_i | \mu_i \sim N(\mu_i, \sigma^2)$$

- The **linear predictor** is:

$$\eta_i = \beta_0 + \beta_1 x_i \quad (\text{Note the absence of an error term.})$$

- The **link function** is the “identity link”:

$$\eta_i = \mu_i \quad (\eta_i \text{ is set equal to the expected value of the response function})$$

Putting it all together, we have:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad e_i \sim N(0, \sigma^2)$$

Generalized linear modeling (GLM)

GLM separates the response distribution from the link function.

So, GLM extends standard MLR in two ways:

- Choosing a (usually) nonnormal response distribution
- Using a (possibly) nonlinear link function

This is *somewhat* like using a nonlinear transformation of the outcome, but it is built into the model.

Manual transformation of y followed by regression requires us to assume normality for the errors, but the error distribution (and therefore response distribution) may not be so simple after transformation, or the error variance may depend on the mean.

In GLM, this is not a problem.

Logistic regression

Consider the single-level logistic function for binary outcomes.

- The **response distribution** of y is assumed Bernoulli with mean μ (the proportion of “1” responses), where the variance is a simple function of the mean:

$$y_i | \mu_i \sim \text{Ber}(\mu_i) \quad \text{Var}(y_i | \mu_i) = \mu_i(1 - \mu_i)$$

- The **linear predictor** is still:

$$\eta_i = \beta_0 + \beta_1 x_i \quad (\text{Note the absence of an error term.})$$

- The **link function** is the “logit link”:

$$\begin{aligned} \eta_i &= \text{logit}(\mu_i) \\ &= \ln \left(\frac{\mu_i}{1 - \mu_i} \right) \end{aligned}$$

Logistic regression

Why is this a sensible model for a binary y ?

We know y can have only two values: 0 and 1. So it is sensible to model the *probability of responding 0 or 1* as a function of predictors x .

We have already seen why that model cannot be linear; this would imply an unrealistic distribution of predicted values of y . The Bernoulli distribution is a clear and logical alternative; its mean is the probability of 0 vs. 1, and its variance is a simple function of the mean.

$$\eta_i = \text{logit}(\mu_i)$$

$$\beta_0 + \beta_1 x_i = \ln \left(\frac{\mu_i}{1 - \mu_i} \right)$$

this is a “logit” or “log odds”

- numerator = probability of a “1” response
- denominator = probability of a “0” response
- ratio is the “odds”
- log of the ratio is thus the “log odds”

The logit is negative when $\text{prob}(1) < .5$ and positive when $\text{prob}(1) > .5$. The logit is “linear in the parameters,” so the regression is linear in terms of the logit, not in terms of the probabilities.

Logistic regression

To link y explicitly to its predictors x requires using an inverse link.

$$\text{If } \eta_i = \ln \left(\frac{\mu_i}{1 - \mu_i} \right), \text{ then } \mu_i = \frac{1}{1 + \exp(-\eta_i)}.$$

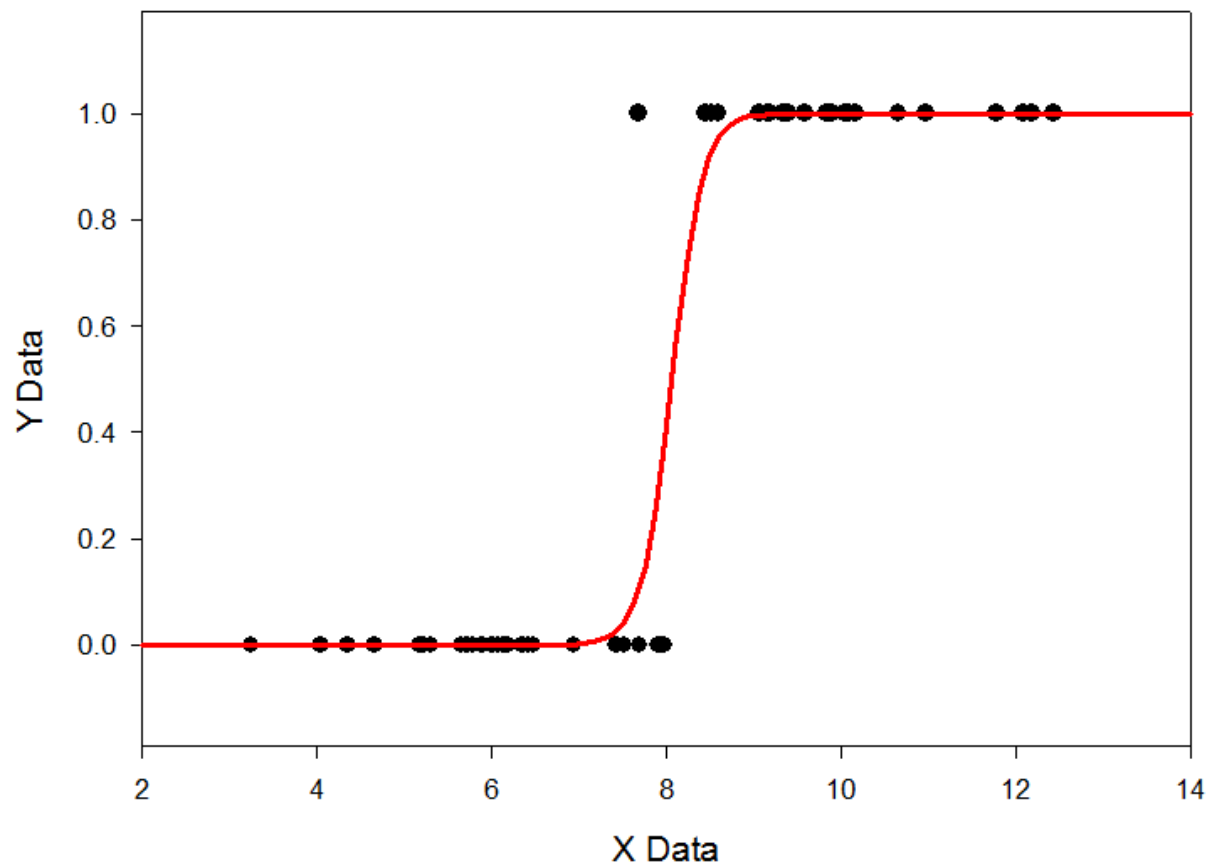
Therefore,

$$\begin{aligned} \mu_i &= \frac{1}{1 + \exp(-\beta_0 - \beta_1 x_i)} \\ &= \frac{1 / \exp(-\beta_0 - \beta_1 x_i)}{1 / \exp(-\beta_0 - \beta_1 x_i) + \exp(-\beta_0 - \beta_1 x_i) / \exp(-\beta_0 - \beta_1 x_i)} \\ &= \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \quad \dots \text{the logistic function.} \end{aligned}$$

Logistic regression

$$\mu_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

This inverse link function has an upper asymptote of 1, and a lower asymptote of 0, which is desirable for binary (0,1) outcomes.



For these data,

$$\beta_0 = -46.013$$

$$\beta_1 = 5.707$$

Logistic regression: Interpretation

In logistic regression, the parameters can be converted into **odds ratios**.

$$\eta_i = \ln\left(\frac{\mu_i}{1 - \mu_i}\right)$$

$$\exp(\eta_i) = \frac{\mu_i}{1 - \mu_i}$$

an "odds"



$$\exp(\beta_0 + \beta_1 x_i) = \frac{\mu_i}{1 - \mu_i}$$

$$\exp(\beta_0) \exp(\beta_1 x_i) = \frac{\mu_i}{1 - \mu_i}$$

$$\exp(\beta_0) \exp(\beta_1)^{x_i} = \frac{\mu_i}{1 - \mu_i}$$

Logistic regression: Interpretation

$$\exp(\beta_0) \exp(\beta_1)^{x_i} = \frac{\mu_i}{1 - \mu_i}$$

Therefore, $\exp(\beta_0)$ is the odds when $x = 0$, and β_0 is the log odds when $x = 0$.

Also, $\exp(\beta_1)$ is the *factor* by which the odds increases per unit increase in x . Consider moving from $x = a$ to $x = a + 1$:

$$\frac{\exp(\beta_0) \exp(\beta_1)^{a+1}}{\exp(\beta_0) \exp(\beta_1)^a} = \frac{\cancel{\exp(\beta_0)} \exp(\beta_1)^a \exp(\beta_1)^1}{\cancel{\exp(\beta_0)} \exp(\beta_1)^a} = \exp(\beta_1)$$

...so $\exp(\beta_1)$ is the ratio of the odds *after* moving 1 unit on x to the odds *before* moving 1 unit on x , or the **odds ratio**. An OR of 1 indicates no effect of x on the odds of obtaining 0 vs. 1. OR > 1 indicates an increasing probability of obtaining a 1 (vs. 0) as x increases, and OR < 1 indicates the opposite.

β_1 is also the change in the log odds for a unit change in x .

GLM: Recap

GLM has three components:

- The **response distribution**
- The **linear predictor**
- The **link function**

Logistic regression is just one example of GLM, with one choice of response distribution, linear predictor, and link function.

Other examples include:

- Multinomial regression (for unordered categorical outcomes)
- Poisson regression (for count outcomes)
- Cox regression (for survival / time-to-event outcomes)
- Tobit regression (for censored outcomes)

...etc.

GLM: Probit regression

In principle, many response distributions can be used with any link function, and many different link functions can be used with any response distribution.

For example, the logit link is not the only link function used with binary data.

Another common link function is the **probit link**:

$$\eta_i = \Phi^{-1}(\mu_i)$$

where $\Phi^{-1}(\mu_i)$ is the inverse Gaussian (or inverse normal) function.

GLM: Probit regression

$$\eta_i = \Phi^{-1}(\mu_i)$$

This function turns probabilities of responding 1 vs. 0 (think p -values under a normal curve) into z -scores, which range continuously from negative infinity to positive infinity.

β_1 from a probit model can be interpreted almost like a Cohen's d .

Probit regression is the same as assuming there is a continuous normal variable (y^*) underlying our binary y , and β_1 is a standardized mean change in y^* per unit change in x .

Logistic regression tends to be more popular because its coefficients can be transformed into odds ratios (ORs).

There are only slight differences between logistic and probit regression weights. The biggest difference is simply due to the errors having different variances.

$$\boldsymbol{\beta}_L \approx \sqrt{\pi^2 / 3} (\boldsymbol{\beta}_P)$$

Generalized linear mixed model (GLMM)

Just as in the continuous outcomes case, if there is clustering in the case of discrete outcomes, that clustering must be taken into account in the model or we risk biased standard errors and violation of independence.

Begin with a simple multilevel model with a random intercept and random slope:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + e_{ij} \quad e_{ij} \sim N(0, \sigma_e^2)$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix} \right)$$

Generalized linear mixed model (GLMM)

We can recast this as a special case of the **generalized linear mixed model (GLMM)**.

Response distribution:

$$y_{ij} | \mu_{ij} \sim N(\mu_{ij}, \sigma_e^2)$$

only the level-1 variance goes here

Linear predictor:

the multilevel part is
considered part of the linear predictor

$$\eta_{ij} = \beta_{0j} + \beta_{1j}x_{1ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ & \tau_{11} \end{bmatrix} \right)$$

Link function:

$$\eta_{ij} = \mu_{ij}$$

identity link

Multilevel logistic model

For binary outcomes, specify the response distribution as Bernoulli, with a logit link.

Response distribution:

$$y_{ij} \mid \mu_{ij} \sim \text{Ber}(\mu_{ij})$$

Linear predictor:

the multilevel part is
considered part of the linear predictor

$$\eta_{ij} = \beta_{0j} + \beta_{1j}x_{1ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ & \tau_{11} \end{bmatrix} \right)$$

Link function:

$$\eta_{ij} = \text{logit}(\mu_{ij})$$

logit link

Multilevel logistic model

Single-level logistic regression:

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_i$$

Multilevel logistic regression (adds a random intercept):

$$\ln\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \gamma_{00} + \gamma_{01} x_{ij} + u_{0j}$$

Multilevel logistic model: ICC

We can think of the discrete outcome as reflecting an underlying “latent” variable y^* .

When the logistic formulation is used, an ICC for binary outcomes can be computed as (Snijders & Bosker, 2012, p. 305):

$$\text{ICC}^* = \frac{\tau_{00}}{\tau_{00} + \pi^2 / 3}$$

Interpretable as the proportion of variance that is between clusters on the latent y^* , not on the binary outcome y .

If it is not sensible to think of a latent continuous variable underlying the binary y , then this ICC is not sensible.

In a probit model, the formula is (Hedeker & Gibbons, 2006, p. 158):

$$\text{ICC}^* = \frac{\tau_{00}}{\tau_{00} + 1}$$

Generalized linear mixed model (GLMM)

GLMM is just like GLM, except:

- We add a j subscript to keep track of cluster membership.
- We model level-2 residuals in the linear predictor equation.

That is, we model the y scores as being independent, and its mean is conditional not only on covariates x (and maybe w), but also conditional on cluster membership.

We can use a variety of link functions and response distributions, depending on the nature of the outcome variable y .

Three levels are handled by simply adding level-3 equations to the linear predictor part of the model.

Common response distributions and link functions

| <u>Outcome type</u> | <u>Response dist.</u> | <u>Link function</u> |
|---------------------|-----------------------|----------------------|
| Continuous | Normal | Identity |
| Binary | Bernoulli | Logit |
| Count | Poisson | Log |
| Ordinal | Multinomial | Cumulative logit |
| Nominal | Multinomial | Multinomial logit |

Where to read more about GLMM

Sources for additional information on GLMM:

Raudenbush and Bryk (2002), chapter 10

Binary, count, ordinal, multinomial

Hox (2010), chapters 6-8

Binary, count, ordinal, survival / event history

Snijders and Bosker (2012), chapter 17

Binary, count, ordinal, survival / event history

Hedeker and Gibbons (2006), chapters 9-12

Binary, count, ordinal, multinomial, survival / event history

Twisk (2006), chapters 4 and 9

Binary, count, multinomial, survival / event history

Heck, Thomas, and Tabata (2010)

Binary, count, ordinal, multinomial

Other sources

Molenberghs and Verbeke (2005)

Vonesh and Chinchilli (1997)

Fitzmaurice, Laird, and Ware (2011; 2nd ed.)

McCulloch, Searle, and Neuhaus (2008)

Dobson and Barnett (2008)

Smithson and Merkle (2014)

McCullagh and Nelder (1989; 2nd ed.)

Example 1: Count outcome

Count data are discrete, bounded from below by 0, usually skewed, and often zero-inflated.

Therefore, standard linear models should not be used.

Example 1: Count outcome

The standard GLM for counts uses a Poisson response distribution with a log link.

- The **response distribution** of y is assumed Poisson with mean μ , where the variance is equal to the mean:

$$y_i | \mu_i \sim \text{Poisson}(\mu_i) \quad \text{Var}(y_i | \mu_i) = \mu_i$$

- The **linear predictor** is:

$$\eta_i = \beta_0 + \beta_1 x_i \quad (\text{Note the absence of an error term})$$

- The **link function** is the “log link”:

$$\eta_i = \ln(\mu_i)$$

Example 1: Count outcome

The data were collected from 9956 ninth grade students nested in 44 high schools (Heck, Thomas, & Tabata, 2012).

The outcome of interest is the number of classes each student fails during the 9th grade. Counts range from 0 to 4, heavily weighted toward 0.

We are interested in assessing potential determinants of course failure.

Example 1: Count outcome

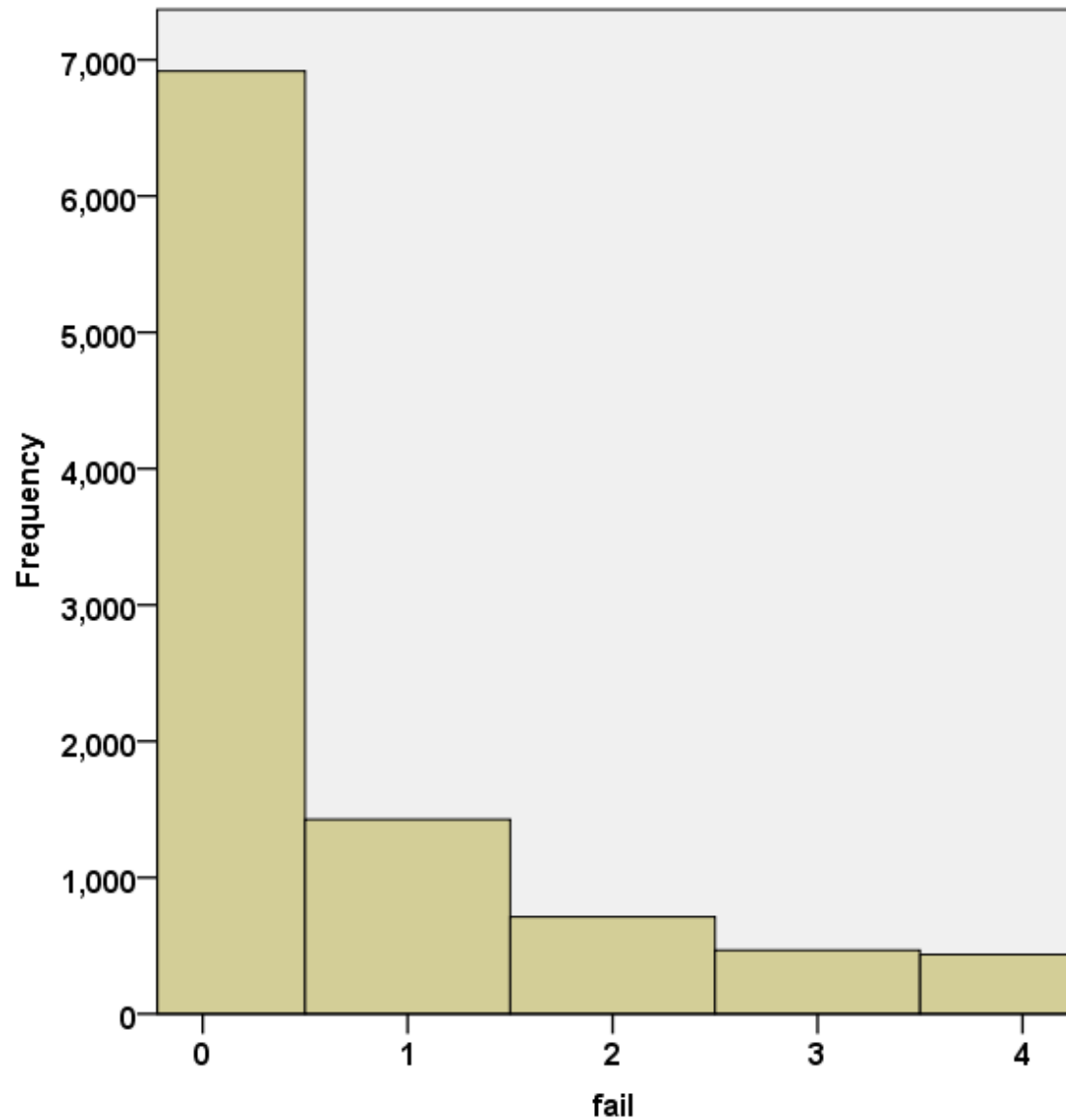
The data:

69.5% of the students failed 0 classes.

30.5% failed at least 1 class.

4.4% failed all 4 classes.

We will use a Poisson distribution, which is a common way to describe counts with rare events.

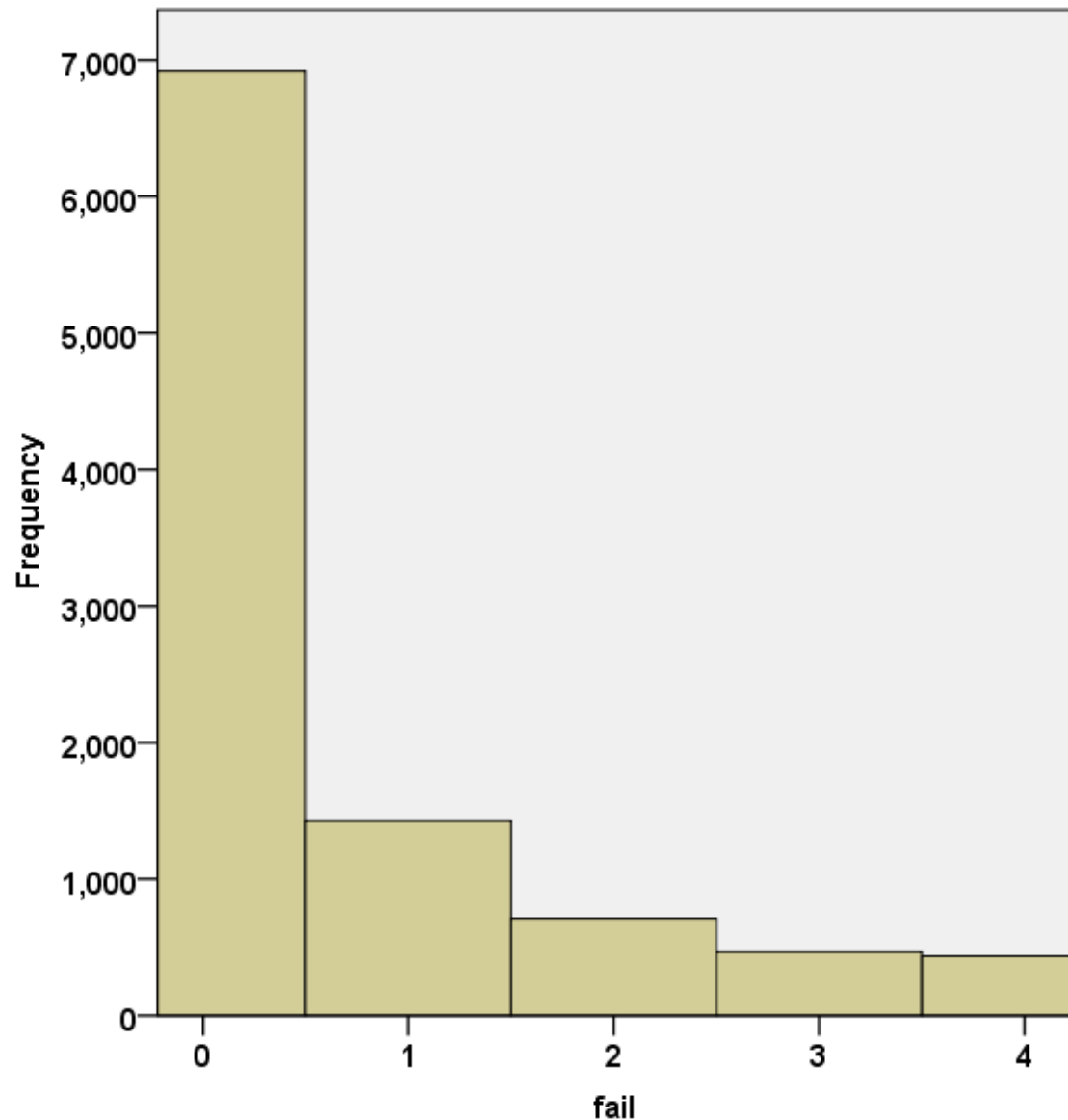


Example 1: Count outcome

In the Poisson distribution, the mean equals the variance.

That is, higher maximum counts tend to pull the mean and the variance up.

But variances increase “faster” than means. If the variance is notably larger than the mean, **overdispersion** is present and robust standard errors are advised.



Example 1: Count outcome

First, a single-level model that ignores clusters.

Four predictors:

- Student SES (0 = average or high; 1 = low)
- Student gender (0 = female; 1 = male)
- Student age at beginning of 9th grade (years)
- Student math achievement in 8th grade (100 – 500)

Example 1: Count outcome

The first model (no predictors) will describe the log of the counts:

Response distribution:

$$y_i | \mu_i \sim \text{Poisson}(\mu_i) \quad \text{Var}(y_i | \mu_i) = \mu_i$$

Linear predictor:

$$\eta_i = \beta_0$$

Link function:

$$\eta_i = \ln(\mu_i)$$

Therefore, the inverse link is $\mu_i = \exp(\eta_i) = e^{\eta_i}$.

Example 1: Count outcome

Parameter Estimates

| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | | Exp(B) | 95% Wald Confidence Interval for Exp(B) | |
|-------------|----------------|------------|------------------------------|-------|-----------------|----|------|--------|---|-------|
| | | | Lower | Upper | Wald Chi-Square | df | Sig. | | Lower | Upper |
| (Intercept) | -.508 | .0129 | -.533 | -.482 | 1544.799 | 1 | .000 | .602 | .587 | .617 |
| (Scale) | 1 ^a | | | | | | | | | |

Dependent Variable: fail

Model: (Intercept)

a. Fixed at the displayed value.

$\beta_0 = -.508$ is the natural log of the expected count of course failures in the 9th grade. Exponentiating this yields the expected count itself, .602.

We can solve this for the expected probability of failing a given number of courses, e.g.:

$$p(y = c) = \frac{e^{-\mu_i} \mu_i^c}{c!} \quad p(y = 4) = \frac{e^{-.602} .602^4}{4!} = .003$$

Example 1: Count outcome

In fact...

$$p(y = 0) = \frac{e^{-.602} \cdot .602^0}{0!} = .548$$

$$p(y = 1) = \frac{e^{-.602} \cdot .602^1}{1!} = .330$$

$$p(y = 2) = \frac{e^{-.602} \cdot .602^2}{2!} = .099$$

$$p(y = 3) = \frac{e^{-.602} \cdot .602^3}{3!} = .020$$

$$p(y = 4) = \frac{e^{-.602} \cdot .602^4}{4!} = .003$$

Example 1: Count outcome

What we *really* want to do is find background factors that help explain the distribution of course failure counts.

For this, augment the linear predictor model from:

$$\eta_i = \beta_0$$

to:

$$\eta_i = \beta_0 + \beta_1 \text{lowses} + \beta_2 \text{male} + \beta_3 \text{math} + \beta_4 \text{age}$$

Example 1: Count outcome

Parameter Estimates

| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | | Exp(B) | 95% Wald Confidence Interval for Exp(B) | |
|-------------|----------------|------------|------------------------------|--------|-----------------|----|------|--------|---|-------|
| | | | Lower | Upper | Wald Chi-Square | df | Sig. | | Lower | Upper |
| (Intercept) | -2.473 | .3696 | -3.197 | -1.748 | 44.772 | 1 | .000 | .084 | .041 | .174 |
| lowses | .344 | .0269 | .292 | .397 | 164.383 | 1 | .000 | 1.411 | 1.339 | 1.487 |
| male | .168 | .0262 | .116 | .219 | 40.810 | 1 | .000 | 1.182 | 1.123 | 1.245 |
| math | -.007 | .0002 | -.007 | -.006 | 934.255 | 1 | .000 | .993 | .993 | .994 |
| age | .242 | .0267 | .190 | .295 | 82.620 | 1 | .000 | 1.274 | 1.209 | 1.343 |
| (Scale) | 1 ^a | | | | | | | | | |

Dependent Variable: fail

Model: (Intercept), lowsese, male, math, age

a. Fixed at the displayed value.

$\beta_0 = -2.473$ is the natural log of the expected count of course failures in the 9th grade, when all the predictors = 0. Exponentiating this yields the expected count itself, .084.

This is not particularly meaningful on its own. We may want to center age and math.

Example 1: Count outcome

Parameter Estimates

| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | | Exp(B) | 95% Wald Confidence Interval for Exp(B) | |
|-------------|----------------|------------|------------------------------|-------|-----------------|----|------|--------|---|-------|
| | | | Lower | Upper | Wald Chi-Square | df | Sig. | | Lower | Upper |
| (Intercept) | -.879 | .0246 | -.927 | -.831 | 1277.482 | 1 | .000 | .415 | .396 | .436 |
| lowses | .344 | .0269 | .292 | .397 | 164.383 | 1 | .000 | 1.411 | 1.339 | 1.487 |
| male | .168 | .0262 | .116 | .219 | 40.810 | 1 | .000 | 1.182 | 1.123 | 1.245 |
| gmmath | -.007 | .0002 | -.007 | -.006 | 934.255 | 1 | .000 | .993 | .993 | .994 |
| gmage | .242 | .0267 | .190 | .295 | 82.620 | 1 | .000 | 1.274 | 1.209 | 1.343 |
| (Scale) | 1 ^a | | | | | | | | | |

Dependent Variable: fail

Model: (Intercept), lowsese, male, gmmath, gmage

a. Fixed at the displayed value.

Only the intercept changed.

$\beta_0 = -.879$ is the natural log of the expected count of course failures in the 9th grade for girls, average/high SES, mean math, and mean age. Exponentiating this yields the expected count itself, .415.

Example 1: Count outcome

Parameter Estimates

| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | | Exp(B) | 95% Wald Confidence Interval for Exp(B) | |
|-------------|----------------|------------|------------------------------|-------|-----------------|----|------|--------|---|-------|
| | | | Lower | Upper | Wald Chi-Square | df | Sig. | | Lower | Upper |
| (Intercept) | -.879 | .0246 | -.927 | -.831 | 1277.482 | 1 | .000 | .415 | .396 | .436 |
| lowses | .344 | .0269 | .292 | .397 | 164.383 | 1 | .000 | 1.411 | 1.339 | 1.487 |
| male | .168 | .0262 | .116 | .219 | 40.810 | 1 | .000 | 1.182 | 1.123 | 1.245 |
| gmmath | -.007 | .0002 | -.007 | -.006 | 934.255 | 1 | .000 | .993 | .993 | .994 |
| gmage | .242 | .0267 | .190 | .295 | 82.620 | 1 | .000 | 1.274 | 1.209 | 1.343 |
| (Scale) | 1 ^a | | | | | | | | | |

Dependent Variable: fail

Model: (Intercept), lowsese, male, gmmath, gmage

a. Fixed at the displayed value.

Each slope is the amount by which a 1-unit change in the predictor changes the log of the expected failure rate (holding other predictors constant). All are significant.

Plugging in particular values of the predictors will give us logs of expected counts, so we merely have to (a) provide predictors values and (b) exponentiate the result to obtain the expected failure count at any combination of predictor values.

Example 1: Count outcome

Parameter Estimates

| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | | Exp(B) | 95% Wald Confidence Interval for Exp(B) | |
|-------------|----------------|------------|------------------------------|-------|-----------------|----|------|--------|---|-------|
| | | | Lower | Upper | Wald Chi-Square | df | Sig. | | Lower | Upper |
| (Intercept) | -.879 | .0246 | -.927 | -.831 | 1277.482 | 1 | .000 | .415 | .396 | .436 |
| lowses | .344 | .0269 | .292 | .397 | 164.383 | 1 | .000 | 1.411 | 1.339 | 1.487 |
| male | .168 | .0262 | .116 | .219 | 40.810 | 1 | .000 | 1.182 | 1.123 | 1.245 |
| gmmath | -.007 | .0002 | -.007 | -.006 | 934.255 | 1 | .000 | .993 | .993 | .994 |
| gmage | .242 | .0267 | .190 | .295 | 82.620 | 1 | .000 | 1.274 | 1.209 | 1.343 |
| (Scale) | 1 ^a | | | | | | | | | |

Dependent Variable: fail

Model: (Intercept), lowsese, male, gmmath, gmage

a. Fixed at the displayed value.

Expected count for females who are average/high SES, avg. math, and avg. age:

$$\exp(-.879) = .415$$

Expected count for males who are average/high SES, avg. math, and avg. age:

$$\exp(-.879 + .168) = \exp(-.879) \exp(.168) = .491$$

Example 1: Count outcome

Parameter Estimates

| Parameter | B | Std. Error | 95% Wald Confidence Interval | | Hypothesis Test | | | Exp(B) | 95% Wald Confidence Interval for Exp(B) | |
|-------------|----------------|------------|------------------------------|-------|-----------------|----|------|--------|---|-------|
| | | | Lower | Upper | Wald Chi-Square | df | Sig. | | Lower | Upper |
| (Intercept) | -.879 | .0246 | -.927 | -.831 | 1277.482 | 1 | .000 | .415 | .396 | .436 |
| lowses | .344 | .0269 | .292 | .397 | 164.383 | 1 | .000 | 1.411 | 1.339 | 1.487 |
| male | .168 | .0262 | .116 | .219 | 40.810 | 1 | .000 | 1.182 | 1.123 | 1.245 |
| gmmath | -.007 | .0002 | -.007 | -.006 | 934.255 | 1 | .000 | .993 | .993 | .994 |
| gmage | .242 | .0267 | .190 | .295 | 82.620 | 1 | .000 | 1.274 | 1.209 | 1.343 |
| (Scale) | 1 ^a | | | | | | | | | |

Dependent Variable: fail

Model: (Intercept), lowses, male, gmmath, gmage

a. Fixed at the displayed value.

The ratio of these rates, $.491/.415 = 1.182$, is the Exp(B) value reported for “male” in the output. That is, males fail courses at 1.182 times the rate that females do, holding constant the other predictors.

Example 1: Count outcome

This analysis did not account for nesting. Using GLMM...

Response distribution:

$$y_{ij} \mid \mu_{ij} \sim \text{Poisson}(\mu_{ij}) \quad \text{Var}(y_{ij} \mid \mu_{ij}) = \mu_{ij}$$

Linear predictor:

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \beta_{1j}x_{1ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \quad \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \text{MVN} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix} \right)$$

Link function:

$$\eta_{ij} = \ln(\mu_{ij})$$

Example 1: Count outcome

In SPSS, multilevel GLM is fit using GENLIMIXED.

The same model we just fit can be specified using GENLIMIXED with no level-2 error terms (i.e., fixed intercept and slopes—no /RANDOM statement).

GENLIMIXED

```
/DATA_STRUCTURE SUBJECTS=nschcode  
/FIELDS TARGET=fail TRIALS=NONE OFFSET=NONE  
/TARGET_OPTIONS DISTRIBUTION=POISSON LINK=LOG  
/FIXED EFFECTS=lowses male gmmath gmage USE_INTERCEPT=TRUE  
/BUILD_OPTIONS TARGET_CATEGORY_ORDER=ASCENDING  
  INPUTS_CATEGORY_ORDER=ASCENDING MAX_ITERATIONS=100  
  CONFIDENCE_LEVEL=95 DF_METHOD=RESIDUAL COVB=ROBUST  
/EMMEANS_OPTIONS SCALE=ORIGINAL PADJUST=LSD.
```

Example 1: Count outcome

The default output type contains a mix of tabular text and graphics. Here I focus on the text output for fixed coefficients:

Fixed Coefficients^a

| Model Term | Coefficient | Std. Error | t | Sig. | 95% Confidence Interval | | Exp (Coefficient) | 95% Confidence Interval for Exp (Coefficient) | |
|------------|-------------|------------|---------|------|-------------------------|-------|----------------------|--|-------|
| | | | | | Lower | Upper | | Lower | Upper |
| Intercept | -.879 | .0330 | -26.674 | .000 | -.944 | -.815 | .415 | .389 | .443 |
| lowses | .344 | .0372 | 9.248 | .000 | .271 | .417 | 1.411 | 1.312 | 1.518 |
| male | .168 | .0358 | 4.680 | .000 | .097 | .238 | 1.182 | 1.102 | 1.269 |
| gmmath | -.007 | .0003 | -24.733 | .000 | -.007 | -.006 | .993 | .993 | .994 |
| gimage | .242 | .0375 | 6.459 | .000 | .169 | .316 | 1.274 | 1.184 | 1.372 |

Probability distribution: Poisson

Link function: Log

a. Target: fail

...and the only random effect, the intercept variance:

Residual Effect

| Residual Effect | Estimate | Std. Error | Z | Sig. | 95% Confidence Interval | |
|-----------------|----------|------------|---|------|-------------------------|-------|
| | | | | | Lower | Upper |
| Variance | 1.806 | .000 | . | . | 1.806 | 1.806 |

Covariance Structure: Scaled Identity

Subject Specification: (None)

Example 1: Count outcome

Allowing both the intercept and the SES slope to be random...

GENLINMIXED

```
/DATA_STRUCTURE SUBJECTS=nschcode  
/FIELDS TARGET=fail TRIALS=NONE OFFSET=NONE  
/TARGET_OPTIONS DISTRIBUTION=POISSON LINK=LOG  
/FIXED EFFECTS=lowses male gmmath gmage USE_INTERCEPT=TRUE  
/RANDOM EFFECTS=lowses USE_INTERCEPT=TRUE SUBJECTS=nschcode  
COVARIANCE_TYPE=UNSTRUCTURED  
/BUILD_OPTIONS TARGET_CATEGORY_ORDER=ASCENDING  
INPUTS_CATEGORY_ORDER=ASCENDING MAX_ITERATIONS=100  
CONFIDENCE_LEVEL=95 DF_METHOD=RESIDUAL COVB=ROBUST  
/EMMEANS_OPTIONS SCALE=ORIGINAL PADJUST=LSD.
```


Example 1: Count outcome

The output in tabular format (fixed effects only):

Fixed Coefficients^a

| Model Term | Coefficient | Std. Error | t | Sig. | 95% Confidence Interval | | Exp (Coefficient) | 95% Confidence Interval for Exp (Coefficient) | |
|------------|-------------|------------|---------|------|-------------------------|-------|----------------------|--|-------|
| | | | | | Lower | Upper | | Lower | Upper |
| Intercept | -1.084 | .0929 | -11.673 | .000 | -1.266 | -.902 | .338 | .282 | .406 |
| lowses | .346 | .0516 | 6.708 | .000 | .245 | .447 | 1.414 | 1.278 | 1.564 |
| male | .163 | .0322 | 5.068 | .000 | .100 | .226 | 1.177 | 1.105 | 1.254 |
| gmmath | -.007 | .0004 | -16.624 | .000 | -.008 | -.007 | .993 | .992 | .993 |
| gimage | .197 | .0450 | 4.376 | .000 | .109 | .285 | 1.218 | 1.115 | 1.330 |

Probability distribution: Poisson

Link function: Log

a. Target: fail

Example 1: Count outcome

Random effect (co)variances:

| Residual Effect | | | | | | |
|-----------------|----------|------------|---|------|-------------------------|-------|
| Residual Effect | Estimate | Std. Error | Z | Sig. | 95% Confidence Interval | |
| | | | | | Lower | Upper |
| Variance | 1.000 | . | . | . | . | . |

Covariance Structure: Scaled Identity
Subject Specification: (None)

| Random Effect | | | | | | | |
|--------------------------|----------|------------|--------|------|-------------------------|-------|--|
| Random Effect Covariance | Estimate | Std. Error | Z | Sig. | 95% Confidence Interval | | |
| | | | | | Lower | Upper | |
| UN (1,1) | .295 | .081 | 3.667 | .000 | .173 | .504 | |
| UN (2,1) | -.067 | .037 | -1.806 | .071 | -.140 | .006 | |
| UN (2,2) | .066 | .025 | 2.654 | .008 | .032 | .139 | |

Covariance Structure: Unstructured
Subject Specification: nschode

Tau matrix

Example 2: Binary outcome

The data were collected from 6528 students nested in 122 schools (Heck, Thomas, & Tabata, 2012).

The outcome of interest is **reading proficiency** (0 = not proficient; 1 = proficient).

We are interested in assessing potential determinants of reading proficiency. These might include:

lowses: socioeconomic status (0,1)

female: being female (0,1)

Example 2: Binary outcome

Consider the logistic function for binary outcomes.

- The **response distribution** of y is assumed Bernoulli with mean μ (the proportion of “1” responses), where the variance is a function of the mean:

$$y_i | \mu_i \sim \text{Ber}(\mu_i) \quad \text{Var}(y_i | \mu_i) = \mu_i(1 - \mu_i)$$

- The **linear predictor** is still:

$$\eta_i = \beta_0 + \beta_1 x_i \quad (\text{Note the absence of an error term.})$$

- The **link function** is the “logit link”:

$$\begin{aligned} \eta_i &= \text{logit}(\mu_i) \\ &= \ln\left(\frac{\mu_i}{1 - \mu_i}\right) \end{aligned}$$

Example 2: Binary outcome

Using logistic regression, we find:

Variables in the Equation

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---------------------|----------|--------|------|---------|----|------|--------|
| Step 1 ^a | lowses | -1.248 | .057 | 483.655 | 1 | .000 | .287 |
| | female | .455 | .056 | 65.089 | 1 | .000 | 1.576 |
| | Constant | 1.232 | .049 | 637.588 | 1 | .000 | 3.427 |

a. Variable(s) entered on step 1: lowsese, female.

High reading proficiency is associated with not being low SES and with being female (both slopes are significant).

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = 1.232 - 1.248(\text{lowses}) + .455(\text{female})$$

Example 2: Binary outcome

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = 1.232 - 1.248(\text{lowses}) + .455(\text{female})$$

That is, for high SES students, the log odds of being proficient at reading are, for males and females respectively:

$$\ln\left(\frac{\mu_i}{1-\mu_i}\right) = 1.232 \qquad \ln\left(\frac{\mu_i}{1-\mu_i}\right) = 1.232 + .455$$

It follows that the odds of being proficient are, for males and females respectively,

$$\frac{\mu_i}{1-\mu_i} = e^{1.232} = 3.428 \qquad \frac{\mu_i}{1-\mu_i} = e^{1.232} e^{.455} = 5.403$$

Example 2: Binary outcome

$$\frac{\mu_i}{1 - \mu_i} = e^{1.232} = 3.428$$

$$\frac{\mu_i}{1 - \mu_i} = e^{1.232} e^{.455} = 5.403$$

The odds ratio for being female is:

$$\text{OR} = \frac{5.403}{3.428} = 1.576$$

Variables in the Equation

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---------------------|----------|--------|------|---------|----|------|--------|
| Step 1 ^a | lowses | -1.248 | .057 | 483.655 | 1 | .000 | .287 |
| | female | .455 | .056 | 65.089 | 1 | .000 | 1.576 |
| | Constant | 1.232 | .049 | 637.588 | 1 | .000 | 3.427 |

a. Variable(s) entered on step 1: lows, female.

Example 2: Binary outcome

We can squeeze yet more information from these results...

Probability that a high-SES male
is proficient at reading

$$\frac{\mu_i}{1 - \mu_i} = 3.428$$

$$\mu_i = 3.428 - 3.428\mu_i$$

$$\mu_i + 3.428\mu_i = 3.428$$

$$\mu_i(3.428 + 1) = 3.428$$

$$\mu_i = \frac{3.428}{3.428 + 1}$$

$$\mu_i = \mathbf{.774}$$

Probability that a high-SES female
is proficient at reading

$$\frac{\mu_i}{1 - \mu_i} = 5.403$$

$$\mu_i = 5.403 - 5.403\mu_i$$

$$\mu_i + 5.403\mu_i = 5.403$$

$$\mu_i(5.403 + 1) = 5.403$$

$$\mu_i = \frac{5.403}{5.403 + 1}$$

$$\mu_i = \mathbf{.844}$$

Example 2: Binary outcome

$$\text{OR} = \frac{5.403}{3.428} = 1.576$$

If the odds of being proficient are the same for males and females, $\text{OR} = 1$.

If the odds of being proficient are greater for males, $\text{OR} < 1$.

If the odds of being proficient are greater for females, $\text{OR} > 1$.

Example 2: Binary outcome

Okay, but our observations are nested within schools.

First question: Is there sufficient variability at level-2 (across schools) to warrant GLMM?

Fit a null model!

Example 2: Binary outcome

For binary outcomes, the response distribution is Bernoulli, with a logit link.

Response distribution:

$$y_{ij} \mid \mu_{ij} \sim \text{Ber}(\mu_{ij})$$

Linear predictor:

the multilevel part is
considered part of the linear predictor

$$\eta_{ij} = \beta_{0j} + \beta_{1j}x_{1ij}$$

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim MVN \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ & \tau_{11} \end{bmatrix} \right)$$

Link function:

$$\eta_{ij} = \text{logit}(\mu_{ij})$$

logit link

Example 2: Binary outcome

Using GENLINMIXED...

GENLINMIXED

```
/DATA_STRUCTURE SUBJECTS=schcode  
/FIELDS TARGET=readprof TRIALS=NONE OFFSET=NONE  
/TARGET_OPTIONS DISTRIBUTION=BINOMIAL LINK=LOGIT  
/FIXED USE_INTERCEPT=TRUE  
/RANDOM USE_INTERCEPT=TRUE SUBJECTS=schcode  
  COVARIANCE_TYPE=VARIANCE_COMPONENTS  
/BUILD_OPTIONS TARGET_CATEGORY_ORDER=DESCENDING  
  INPUTS_CATEGORY_ORDER=DESCENDING  
  MAX_ITERATIONS=100 CONFIDENCE_LEVEL=95  
  DF_METHOD=RESIDUAL COVB=ROBUST  
/EMMEANS_OPTIONS SCALE=ORIGINAL PADJUST=LSD.
```

Example 2: Binary outcome

Using GENLINUXED...

Fixed Coefficients^a

| Model Term | Coefficient | Std. Error | t | Sig. | 95% Confidence Interval | | Exp (Coefficient) | 95% Confidence Interval for Exp (Coefficient) | |
|------------|-------------|------------|--------|------|-------------------------|-------|----------------------|--|-------|
| | | | | | Lower | Upper | | Lower | Upper |
| Intercept | .905 | .0796 | 11.366 | .000 | .749 | 1.061 | 2.472 | 2.115 | 2.890 |

Probability distribution: Binomial

Link function: Logit

a. Target: readprof

Random Effect

| Random Effect Covariance | Estimate | Std. Error | Z | Sig. | 95% Confidence Interval | |
|-----------------------------|----------|------------|-------|------|-------------------------|-------|
| | | | | | Lower | Upper |
| Var(Intercept) | .640 | .106 | 6.028 | .000 | .462 | .885 |

Covariance Structure: Variance components

Subject Specification: schcode

The intraclass correlation is therefore $ICC = \frac{.640}{.640 + \pi^2 / 3} = .163$.

So yes, it's worth it to use GLMM.

Example 2: Binary outcome

Using GENLINUX with predictors...

GENLINUX

```
/DATA_STRUCTURE SUBJECTS=schcode  
/FIELDS TARGET=readprof TRIALS=NONE OFFSET=NONE  
/TARGET_OPTIONS DISTRIBUTION=BINOMIAL LINK=LOGIT  
/FIXED EFFECTS=female lowses USE_INTERCEPT=TRUE  
/RANDOM USE_INTERCEPT=TRUE SUBJECTS=schcode  
  COVARIANCE_TYPE=VARIANCE_COMPONENTS  
/BUILD_OPTIONS TARGET_CATEGORY_ORDER=DESCENDING  
  INPUTS_CATEGORY_ORDER=DESCENDING  
  MAX_ITERATIONS=100 CONFIDENCE_LEVEL=95  
  DF_METHOD=RESIDUAL COVB=ROBUST  
/EMMEANS_OPTIONS SCALE=ORIGINAL PADJUST=LSD.
```

Example 2: Binary outcome

Using GENLINMIXED, we obtain fixed effects...

Fixed Coefficients^a

| Model Term | Coefficient | Std. Error | t | Sig. | 95% Confidence Interval | | Exp (Coefficient) | 95% Confidence Interval for Exp (Coefficient) | |
|------------|----------------|------------|---------|------|-------------------------|-------|----------------------|--|-------|
| | | | | | Lower | Upper | | Lower | Upper |
| Intercept | 1.145 | .0857 | 13.352 | .000 | .977 | 1.313 | 3.141 | 2.655 | 3.716 |
| female=1 | .468 | .0585 | 8.006 | .000 | .353 | .583 | 1.597 | 1.424 | 1.791 |
| female=0 | 0 ^b | . | . | . | . | . | . | . | . |
| lowses=1 | -.945 | .0658 | -14.366 | .000 | -1.074 | -.816 | .389 | .342 | .442 |
| lowses=0 | 0 ^b | . | . | . | . | . | . | . | . |

Probability distribution: Binomial

Link function: Logit

a. Target: readprof

b. This coefficient is set to zero because it is redundant.

Example 2: Binary outcome

...and the random residual variance for the intercept:

| Random Effect | | | | | | |
|----------------|----------|------------|-------|------|-------------------------|-------|
| Random Effect | Estimate | Std. Error | Z | Sig. | 95% Confidence Interval | |
| Covariance | | | | | Lower | Upper |
| Var(Intercept) | .384 | .074 | 5.224 | .000 | .264 | .559 |

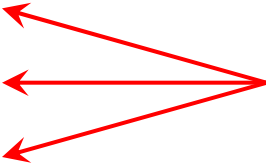
Covariance Structure: Variance components
Subject Specification: schcode

There were no random slopes in this model.

12. Introduction to *Mplus*

Text in, text out

Mplus uses 3 basic files in most analyses:

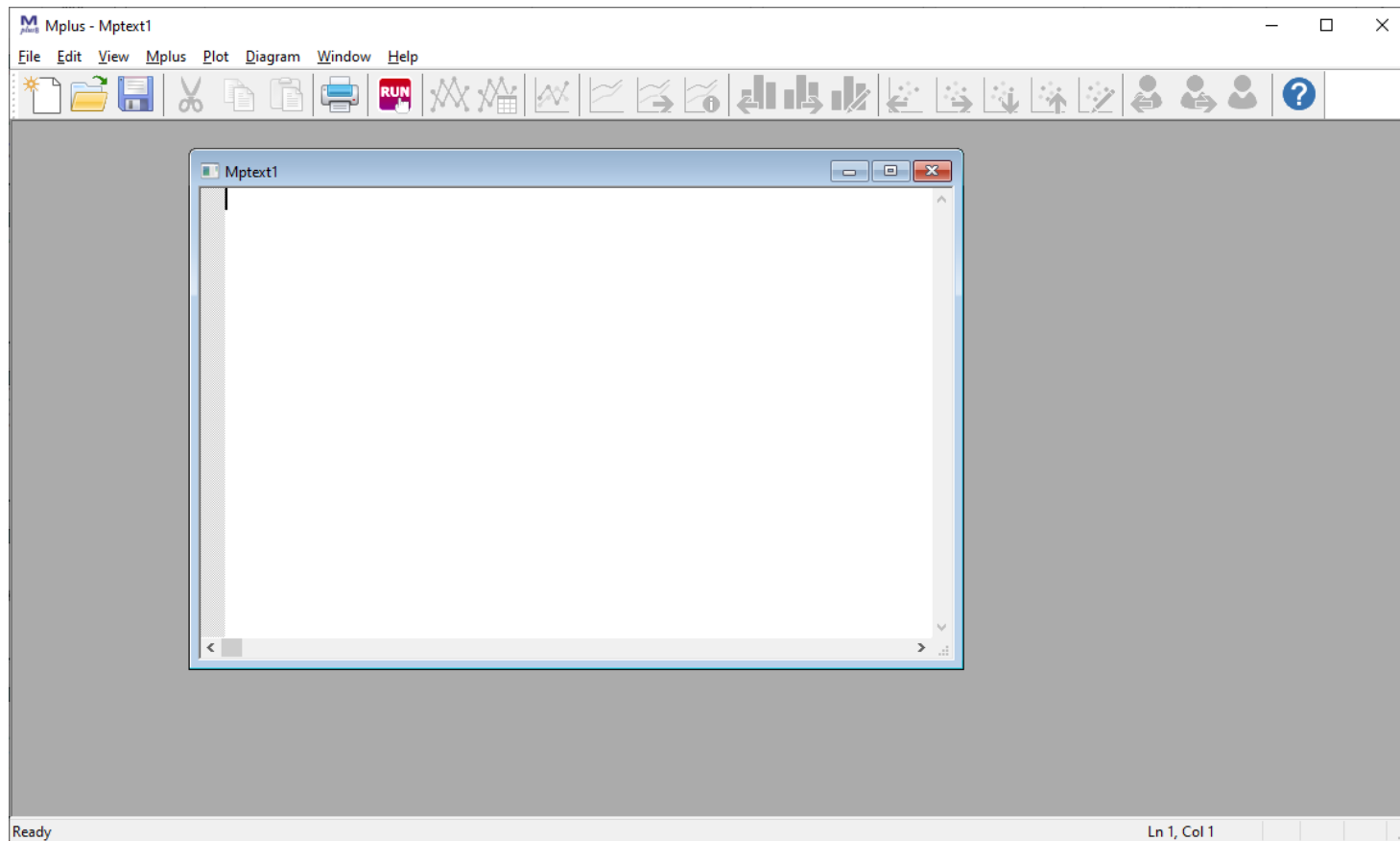
| | | | |
|----------------|------|---|--|
| Data: | .dat |  | All of these are ASCII (text) files |
| Input: | .inp | | |
| Output: | .out | | |

The **data** file contains your data. Missing values need a code (not blanks). MLM requires raw data.

The **input** file is where to tell Mplus about the data, all the options and settings, and what model to fit to the data.

The **output** file contains estimation details and parameter estimates.

Mplus command interface



Example code

Strictly speaking, only some of these are necessary for MLM:

TITLE:

DATA:

VARIABLE:

DEFINE:

ANALYSIS:

MODEL:

OUTPUT:

SAVEDATA:

PLOT:

MONTECARLO:

All these commands are fully documented in the *Mplus User's Guide*, downloadable in its entirety from statmodel.com.

We will cover a *little* more detail here, and then more when needed.

The DATA command

```
DATA: FILE IS mydata.dat;
```

The data file is an ASCII (text) file with numeric data, comma-, tab, or space-delimited.

Limit = 500 variables.

The VARIABLE command

```
VARIABLE: NAMES ARE id x1 z3 insulin cortisol y1-y3;  
USEVARIABLES ARE x1 insulin cortisol y1-y3;  
MISSING ARE ALL (-99);
```

`NAMES ARE` tells Mplus what to call the variables in the data file. Variable names need to be 8 characters or shorter.

`USEVARIABLES ARE` tells Mplus what variables to use in this analysis.

`MISSING ARE` tells Mplus what code you used for missing values (like * or -99). This code can be different for different variables.

The MODEL command

```
MODEL: x1 insulin; x1 WITH insulin;  
cortisol ON x1 insulin;  
glucose BY y1-y3; [glucose];  
glucose ON cortisol x1 insulin;
```

ON regresses a dependent variable (cortisol) on predictors (x_1 and insulin).

BY creates a latent variable (glucose) with three observed indicators ($y_1 - y_3$).

x_1 insulin; estimates variances for these predictors.

x_1 WITH insulin; estimates the covariance of x_1 and insulin.

[glucose]; estimates an intercept for glucose.

The MODEL command

An asterisk (*) means to freely estimate the parameter (usually the default).

```
MODEL: glucose ON cortisol*;
```

An * plus a value frees the parameter and uses the value as a starting value.

```
MODEL: glucose ON cortisol*.26;
```

An “at” sign (@) means to fix the parameter to a particular value.

```
MODEL: glucose ON cortisol@.26;
```

A term in parentheses labels the parameter (useful for constraints).

```
MODEL: glucose ON cortisol(b1);  
        glucose ON insulin(b1);
```

```
MODEL: glucose ON cortisol insulin(b1);
```


Path diagrams

As with SEM, MLM can be illustrated with path diagrams.

A few conventions for diagramming multilevel models have emerged (Curran & Bauer, 2007; Muthén & Asparouhov, 2009; Muthén & Satorra, 1989; Rabe-Hesketh, Skrondal, & Pickles, 2004).

A popular method appears in the *Mplus User's Guide*.

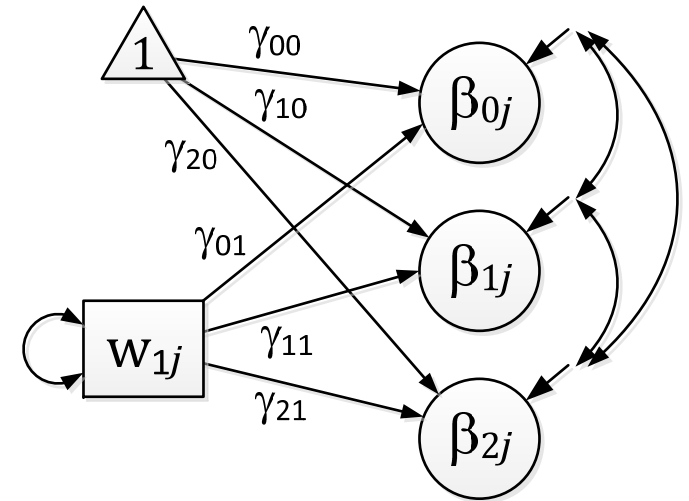
Path diagrams

Level-2 equations:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}w_{1j} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}w_{1j} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}w_{1j} + u_{2j}$$



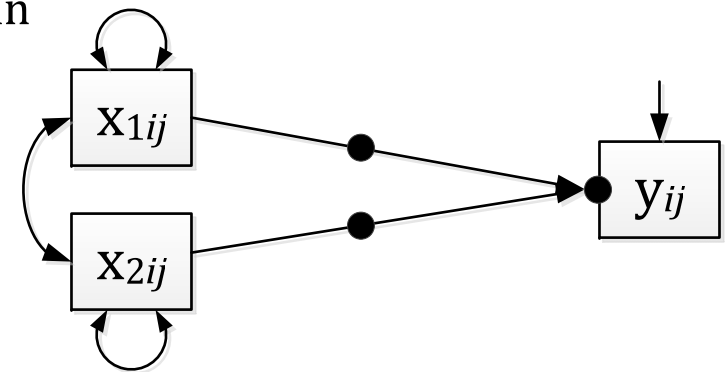
Between

Level-1 equation:

$$y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + e_{ij}$$



Within



Black dots signify random coefficients.

A simple example in Mplus

This MLM example uses the High School and Beyond (HSAB) data.

$N = 7,185$ students, $J = 160$ schools

| | |
|-----------------|-----------------------------|
| schid | School ID |
| minority | 1 = minority, 0 = other |
| female | 1 = female, 0 = male |
| ses | parent socioeconomic status |
| mathach | mathematics achievement |
| size | school enrollment |
| sector | 1 = Catholic, 0 = public |
| meanses | school mean SES* |

*close to, but not the same as, the mean of *observed* SES.

Random effects ANOVA (“random intercepts only” model)

$$\mathit{mathach}_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$
$$u_{0j} \sim N(0, \tau_{00})$$
$$e_{ij} \sim N(0, \sigma^2)$$

A random effects ANOVA model simply splits $\mathit{mathach}_{ij}$ into “within” and “between” components.

Because these components are uncorrelated, the variance of the outcome is also split into two components that add to yield the total variance of $\mathit{mathach}_{ij}$.

$$\text{var}(\mathit{mathach}_{ij}) = \tau_{00} + \sigma^2$$

Random effects ANOVA (“random intercepts only” model)

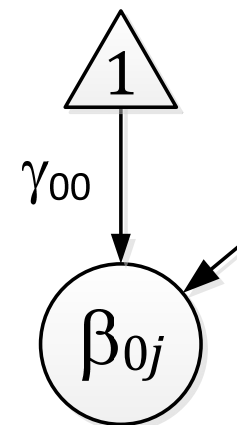
The principle reasons to carry out a random effects ANOVA are to estimate the variance components, estimate the ICC, and to serve as a baseline “null” model for adding predictors.

Random effects ANOVA (“random intercepts only” model)

```
TITLE: hsab random effects anova;  
DATA: FILE IS hsab.dat;  
VARIABLE:  
NAMES ARE school minority female  
ses mathach size sector meanses;  
USEVARIABLES ARE mathach;  
CLUSTER IS school;  
ANALYSIS: TYPE IS TWOLEVEL;  
MODEL:
```

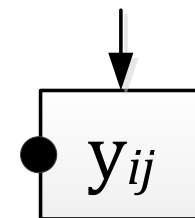
```
%WITHIN%  
mathach;
```

```
%BETWEEN%  
mathach; [mathach];
```



Between

Within



Random effects ANOVA (“random intercepts only” model)

| | Estimate | S.E. | Two-Tailed Est./S.E. | P-Value |
|---------------|----------|-------|-------------------------|---------|
| Within Level | | | | |
| Variances | | | | |
| MATHACH | 39.148 | 0.835 | 46.876 | 0.000 |
| Between Level | | | | |
| Means | | | | |
| MATHACH | 12.637 | 0.244 | 51.824 | 0.000 |
| Variances | | | | |
| MATHACH | 8.562 | 1.057 | 8.101 | 0.000 |

Random intercept; level-1 and level-2 predictors

$$\mathit{mathach}_{ij} = \beta_{0j} + \beta_{1j}\mathit{ses}_{ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\mathit{sector}_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

$$u_{0j} \sim N(0, \tau_{00})$$

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

This model includes SES as a level-1 predictor and school sector (public vs. Catholic) as a level-2 predictor.

Intercepts are random. τ_{00} now represents residual variance at level-2.

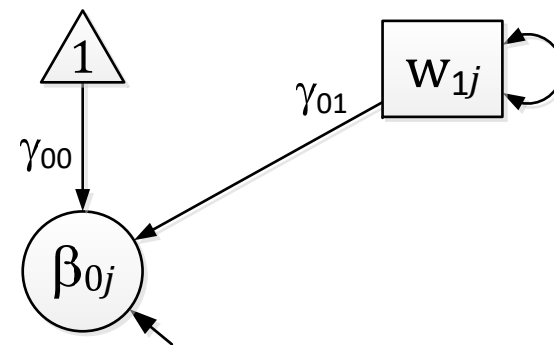
ses_{ij} can potentially explain “within” and “between” variance in $\mathit{mathach}_{ij}$, but sector_j can potentially explain only “between” variance.

Random intercept; level-1 and level-2 predictors

```
TITLE: hsab random intercept, L1 and L2 predictors;  
DATA: FILE IS hsab.dat;  
VARIABLE: NAMES ARE school minority female  
ses mathach size sector meanses;  
USEVARIABLES ARE ses mathach sector;  
CLUSTER IS school;  
WITHIN IS ses; BETWEEN IS sector;  
ANALYSIS: TYPE IS TWOLEVEL;  
MODEL:
```

```
%WITHIN%  
mathach; mathach ON ses;
```

```
%BETWEEN%  
[mathach]; mathach;  
mathach ON sector;
```



Between

Within



Random intercept; level-1 and level-2 predictors

| | Estimate | S.E. | Est./S.E. | Two-Tailed P-Value |
|-------------------------------|----------|-------|-----------|-----------------------|
| Within Level | | | | |
| MATHACH ON SES | 2.376 | 0.128 | 18.589 | 0.000 |
| Residual Variances MATHACH | 37.032 | 0.717 | 51.665 | 0.000 |
| Between Level | | | | |
| MATHACH ON SECTOR | 2.101 | 0.347 | 6.046 | 0.000 |
| Intercepts MATHACH | 11.719 | 0.226 | 51.860 | 0.000 |
| Residual Variances MATHACH | 3.627 | 0.588 | 6.165 | 0.000 |

Random intercept and slope; level-1 and level-2 predictors

$$\begin{aligned} \mathit{mathach}_{ij} &= \beta_{0j} + \beta_{1j} \mathit{ses}_{ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} \mathit{sector}_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \quad \begin{aligned} \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} &\sim \text{MVN} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{pmatrix} \right] \\ e_{ij} &\sim N(0, \sigma^2) \end{aligned}$$

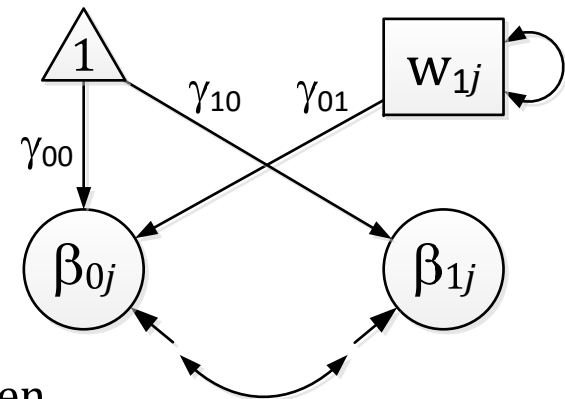
This model includes SES as a level-1 predictor and school sector (public vs. Catholic) as a level-2 predictor.

Intercepts and slopes are now random.

Random intercept and slope; level-1 and level-2 predictors

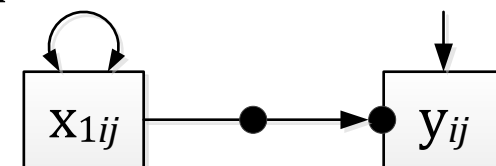
```
TITLE: hsab random intercept and slope, L1 and L2 predictors;  
DATA: FILE IS hsab.dat;  
VARIABLE: NAMES ARE school minority female  
ses mathach size sector meanses;  
USEVARIABLES ARE ses mathach sector;  
CLUSTER IS school;  
WITHIN IS ses; BETWEEN IS sector;  
ANALYSIS: TYPE IS TWOLEVEL RANDOM;  
MODEL:
```

```
%WITHIN%  
mathach; s1 | mathach ON ses;  
  
%BETWEEN%  
[mathach s1];  
mathach s1; mathach WITH s1;  
mathach ON sector;
```



Between

Within



Random intercept and slope; level-1 and level-2 predictors

| | Estimate | S.E. | Two-Tailed Est./S.E. | P-Value |
|----------------------|----------|-------|-------------------------|---------|
| Within Level | | | | |
| Residual Variances | | | | |
| MATHACH | 36.783 | 0.723 | 50.902 | 0.000 |
| Between Level | | | | |
| MATHACH ON SECTOR | 2.530 | 0.423 | 5.979 | 0.000 |
| MATHACH WITH S1 | 0.699 | 0.356 | 1.964 | 0.050 |
| Means | | | | |
| S1 | 2.385 | 0.128 | 18.707 | 0.000 |
| Intercepts | | | | |
| MATHACH | 11.468 | 0.282 | 40.610 | 0.000 |
| Variances | | | | |
| S1 | 0.421 | 0.234 | 1.798 | 0.072 |
| Residual Variances | | | | |
| MATHACH | 3.864 | 0.644 | 5.996 | 0.000 |

$$\mathbf{T} = \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix}$$

$$= \begin{bmatrix} 3.864 & \\ .699 & .421 \end{bmatrix}$$

13. References

Books about Mplus

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2016). *Regression and mediation analysis using Mplus*. Los Angeles: Muthén & Muthén.

Detailed discussion of regression and mediation analysis with Mplus, with many syntax and output files available online at statmodel.com.

Muthén, L. K., & Muthén, B. O. (1998–2016). *Mplus user's guide*. Los Angeles: Muthén & Muthén.

The definitive guide to Mplus capabilities, models, and syntax. Bound versions are available, but the entire guide can be obtained as a .pdf from statmodel.com.

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York: Routledge.

A wonderful starting point for those just learning Mplus. A very useful supplement to the User's Guide.

Geiser, C. (2013). *Data analysis with Mplus*. New York: The Guilford Press.

Excellent introduction to basic and advanced Mplus applications. Includes a good summary of Mplus commands and an Mplus troubleshooting guide.

Books about Mplus

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. Chichester: John Wiley & Sons.

Goes into somewhat more depth on the methods, well-integrated with application with Mplus. Focuses on SEM and mixture models, less on MLM.

Geiser, C., Crayen, C., & Enders, C. (20??). *Advanced multivariate data analysis with Mplus*. Springer VS.

Not out yet; it's on my amazon.com wishlist. Description promises that it will address MSEM, among other topics.

Kelloway, E. K. (2014). *Using Mplus for structural equation modeling: A researcher's guide (2nd ed.)*. Sage Publications.

A user-friendly introduction to Mplus with a focus on SEM.

Online resources

Mplus website (<http://www.statmodel.com>)

- Discussion forum
- Technical support
- *Mplus User's Guide* (free)
- Training handouts
- Web notes / documentation
- Technical appendices
- Videos

SEMNET (<http://www2.gsu.edu/~mkteer/semnet.html>)

Multilevel list (<http://www.bristol.ac.uk/cmm/learning/support/jisc.html>)

quantpsy.org (<http://quantpsy.org/>)

Readings on multilevel modeling

- Bickel, R. (2007). *Multilevel analysis for applied research: It's just regression!* New York: Guilford.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
- Goldstein, H. (2011). *Multilevel statistical models* (4th ed.). Chichester: John Wiley & Sons.
- Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques* (3rd ed.). New York: Routledge.
- Hoffman, L. (2015). *Longitudinal analysis: Modeling within-person fluctuation and change*. New York: Routledge.
- Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, 24, 623-641.
- Hox, J. (2002). *Multilevel analysis: Techniques and applications* (2nd ed.). Mahwah, NJ: Erlbaum.
- Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.
- Leyland, A. H., & Goldstein, H. (2001). *Multilevel modelling of health statistics*. Chichester: Wiley.
- Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.
- Raudenbush, S. W. (1989a). "Centering" predictors in multilevel analysis: Choices and consequences. *Multilevel Modelling Newsletter*, 1, 10-12.
- Raudenbush, S. W. (1989b). A response to Longford and Plewis. *Multilevel Modelling Newsletter*, 1, 8-11.
- Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, 59, 1-17.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.
- Twisk, J. W. R. (2006). *Applied multilevel analysis: A practical guide for medical researchers*. Cambridge: Cambridge University Press.

Readings on generalized linear mixed modeling (GLMM)

| | |
|---|---|
| Raudenbush and Bryk (2002), chapter 10 | Binary, count, ordinal, multinomial |
| Hox (2010), chapters 6-8 | Binary, count, ordinal, survival |
| Snijders and Bosker (2012), chapter 17 | Binary, count, ordinal, survival |
| Hedeker and Gibbons (2006), chapters 9-12 | Binary, count, ordinal, multinomial, survival |
| Twisk (2006), chapters 4 and 9 | Binary, count, multinomial, survival |
| Heck, Thomas, and Tabata (2010) | Binary, count, ordinal, multinomial |

Other book sources

Molenberghs and Verbeke (2005)
Vonesh and Chinchilli (1997)
Fitzmaurice, Laird, and Ware (2011; 2nd ed.)
McCulloch, Searle, and Neuhaus (2008)
Dobson and Barnett (2008)
Smithson and Merkle (2014)
McCullagh and Nelder (1989; 2nd ed.)