Introduction to this Course: Multilevel Models (MLMs) for Educational Data

- Topics:
  - What to expect this semester
  - Types of multilevel data
  - From GLMs to MLMs: The big picture
  - What can MLMs do for you? (a sales pitch)

What To Expect This Semester

- I believe that **everyone is capable** and **can significantly benefit**** from learning more quantitative methods (*especially* multilevel models)!
- **Philosophy:** Focus on accessibility + mastery learning
  - Points for trying, for trying until you get it right, and real use (with revision)
- **Materials:** Unit = lecture + example(s); 6 units planned
  - All at the EDF 9850 external course website (hosted out of Canvas on purpose)
  - **Lecture** slides present concepts—the (wordy) **what** and the **why**
  - **Example** documents: reinforce the concepts and show **how using R software**
  - Suggested **readings after lecture** for a broader perspective and further explanation (no book)
  - New this year—Shiny apps for hands-on visualization of key concepts!



** ***Benefits*** *include but are not limited to: Better research, more authorship opportunities, free travel, and actual money*

Because all of the following are multilevel models…



- Random Effects ANOVA or Repeated Measures ANOVA
  - Occasions are nested within persons after controlling for occasion mean differences
- (Latent) Growth Curve Model (where "Latent" implies SEM software)
  - Btw, most longitudinal MLMs can be equivalently estimated as single-level SEMS
- Within-Person Fluctuation Model (e.g., for EMA or intensive longitudinal data)
  - See also "dynamic" SEM or multilevel SEM (even without measurement models!)
- Clustered/Nested Observations Model (e.g., for students within schools)
  - If followed over time in same group, is "clustered longitudinal model"
- Cross-Classified Models (e.g., teacher "value-added" models)
  - See also subjects crossed with items, raters crossed with targets
- Psychometric Models (e.g., factor analysis, item response theory, SEM)
  - Items modeled as nested in persons (or crossed, as in explanatory IRT)

This Semester's Topics

- Review single-level linear models on your own (interaction terms!)
  - See EDF 9770 (for general) and PSQF 6270 (for general*ized*: non-continuous outcomes)

- Planned units (using continuous outcomes only):
  - 1: Intro to course and two-level models for clustered data (no examples)
  - 2: From empty models to level-2 cluster predictors (1 example)
  - 3: Fixed slopes of level-1 people predictors (1 example)
  - 4: Random slopes and cross-level (people*cluster) interactions (1 example)
  - 5: Longitudinal two-level models (2 examples)
    - See also my prior two-semester longitudinal sequence: PSQF 6271 and PSQF 7375
  - 6: Models for two-level cross-classified designs (3 examples)
  - Interested in other extensions (e.g., generalized) not covered here? Just ask me!

What to Expect Regarding R Software

- **Why R via RStudio? Free** to install on your own device (or use CUAPPS)
  - **Expandable** into newer models and output processing (e.g., computations, tables)

- **It's ok** if you don't know how to use R well or at all!
  - I try to minimize the coding to focus on model components and their interpretation
  - For my intro to R from EDF 9770, watch this Demo Video using Lecture 1 Files
  - Claude code or GPT Codex *can* be helpful (ask it to annotate using comments!!)

- **Your R ≠ my R** (I use =, no tidyverse or pipes; readable >> efficient)
- Although only R is class sponsored, most stats packages do MLMs
  - See PSQF 6272 for similar examples also including STATA and SAS
  - I can also help with SPSS and Mplus (but not Python or JMP)

About the Course Requirements (100 points total)

- **In-class activities:** 5 planned; 2 points each for "reasonable" responses
  - Reinforce lecture concepts during class, working alone or in small groups
  - Shiny apps and practice analysis activities
  - Submit answers via Canvas before next class for automatic credit (even if not in class)

- **Formative assessments (FAs):** 7 planned; 2 points each for "thoughtful" individual responses (even if incorrect)
  - Comprehension check (after watching videos) → structured review during next class
  - e.g., concepts, vocabulary, notation; then story problems, responses to reviewers
  - Mostly brief responses; some multiple-choice questions, submitted via Canvas
  - **USING AI WILL DEFEAT THEIR PURPOSE!!!** (and may cost you points)
    - And sloppy: Responses tend to be long but incomplete, filled with empty words

More About the Course Requirements (/100 total)

- **Homework (HW):** 3 planned for 32 points total
  - Targeted practice conducting analyses and interpreting results (in units 2, 3, and 4)
  - Based directly on examples given (no googling or AI needed)
    - AI *can* be used to help understand the R code or Rstudio interface
  - You will each have a unique dataset (with a common story)
  - **Computation** sections: **Instant feedback, infinite attempts**
  - **Results** sections: **Delayed** feedback, multiple choice (so **single** attempt)
  - Complete "HW 0" over the syllabus for up to **2 points** of extra credit

- Submitted using a **custom online system** (coming soon), but you must also submit your (organized) **syntax and output via Canvas** to receive credit
  - I want to know how you did the work (and learn from you)
  - Future You will thank you for these example files!

More About the Course Requirements (/100 total)

- **Data analytic project (44)** that provides guided practice on real data
  - Individually or in pairs (to be decided by September 23)
  - Culminates in a conference-style zoom presentation on November 12 or 19
    - Can be revised once via video recording if desired
  - **Written components**: plan (5), check-in (10), peer review (4)
    - Instructions, requirements, and rubrics for each will be given later
  - **Must use multilevel analysis, ideally for data you personally care about!!**
    - Preferably clustered data, but longitudinal data could also work
    - I will need to approve all data sources in advance (via the plan document)
    - Can include model extensions not covered in class as needed by your data (in which case I will point you towards the closest examples I can find)

More About the Course Requirements

- Because the course material builds **cumulatively…**
  (and things with hard deadlines tend to get done):
  - Coursework is accepted **up to 2 weeks late** with the following point **penalties** each:
    - In-class activity answers and formative assessments (−1); homework (−3)
    - Project: plan (−2), check-in (−4), presentation (−10), and peer review (−2)
  - **Extensions** can be granted if requested in writing at least 2 weeks in advance
  - Due dates may be pushed later, but never sooner

- Your other jobs:
  - **Ask questions**—preferably in class, but any time is better than none
  - Don't wait until the last minute to start homework, and don't be afraid to **ask for help if you get stuck** on one thing for more than 15 minutes
    - Please email me **screenshots** of your how you created your variables, your model code, and relevant output or error message so I can respond easily

More About the Course Format

- **Attendance**: Strongly recommended but not required
  - You choose each week: in-person as a "**roomer**" or online as a "**zoomer**"… but **please do not attend in-person if you are seriously ill!**
  - I will link to **YouTube-hosted recordings** (audio + screen only) at the EDF 9850 external course website so videos and generated transcript are searchable!

- **Last-minute changes** will be announced by 8 AM on class days
  - I will change to zoom-only if I am sick or there is dangerous weather
  - We will be on zoom for 5+ classes (9/17, 10/1, 11/12, 11/19, 12/3) after this for logistical reasons (needing to be "Dr. Mommy" or student presentations)

- There will be no class on 10/22 (I will be at the SMEP annual meeting) or 11/26 (Thanksgiving); we will not meet during finals week

Multilevel Models (MLMs) for Clustered* Data

- **Clustering = Nesting = Grouping = Hierarchies*
  - Key idea: Outcomes with >1 dimension of sampling simultaneously ("**micro**" units are nested in one or more types of "**macro**" units)
  - Each sampling dimension is considered its own "**level**" → **MLM**
  - MLMs can be used to predict outcomes from two-level (or more-level) sampling designs that result in nested and/or crossed observations

- The term "Multilevel Model" (MLM) has many synonyms:
  - **General Linear Mixed-Effects Models** (Fixed + Random = Mixed)
  - **Random Coefficients Models** (Random effects = latent variables)
  - **Hierarchical Linear Models** (HLM, but not = hierarchical regression)
  - Most MLM software is "univariate" → predict 1 outcome at a time
  - Multivariate MLMs can be estimated as "multilevel structural equation models" to predict 2+ outcomes at once (+ address missing predictors)

Examples of **Nested** Designs

- Examples of **two-level** sampling designs:
  - Clustered: Students (level 1) nested in classes/teachers (level 2)
  - Clustered: Patients (level 1) nested in doctors (level 2)
  - Clustered: Citizens (level 1) nested in countries (level 2)
  - Longitudinal: Occasions (level 1) nested in persons (level 2)

- Examples of **three-level** sampling designs:
  - Students (level 1) nested in classes/teachers (level 2) nested in schools (level 3)
  - Patients (level 1) nested in doctors (level 2) nested in hospitals (level 3)
  - Citizens (level 1) nested in survey years (level 2) nested in countries (level 3)
  - Occasions (level 1) nested in persons (level 2) nested in schools (level 3)
  - Within-day occasions (level 1) nested in days (level 2) nested in persons (level 3)

Examples of **Cross-Classified** Designs

- Examples of **two-level cross-classified** sampling designs:
  - Two kinds of nesting: Students (level 1) nested in both schools (level-2) and neighborhoods (crossed at level 2)
  - Repeated measures: Responses (level 1) nested in both subjects (level 2) and items (crossed at level 2)
  - Reliability assessment: Ratings (level 1) nested in both raters (level 2) and targets (crossed at level 2)
  - Students who change classes over time: occasions (level 1) nested in both students (level 2) and classes (crossed at level 2)

- Example of **three-level cross-classified** sampling designs:
  - Ratings (level 1) nested in both children (level 2) and raters (crossed at level 2); raters are nested within sites (level 3)
  - Responses (level 1) nested in both students (level 2) and items (crossed at level 2); students are nested within schools (level 3)

Statistician-y Labels for Organizing Models

- Outcome type: General (normal) vs. General*ized* (not normal)
- Dimensions of sampling: One (so one variance term per outcome) vs. **Multiple** (so multiple variance terms per outcome) → **OUR WORLD**

- **General Linear Models:** conditionally normal outcome distribution, **fixed effects** (identity link; only one dimension of sampling)
- **General*ized* Linear Models: any conditional outcome distribution**, **fixed**effects through **link functions**, no random effects (only one dimension)
- **General Linear Mixed Models:** conditionally normal outcome distribution, **fixed and random effects**(identity link, but **multiple dimensions** of sampling)
- **General*ized* Linear Mixed Models: any conditional outcome distribution**, **fixed and random effects**through **link functions** (**multiple dimensions**)
  - Same concepts as for general*ized* or mixed separately, but with more complexity in estimation

- "**Linear**" → fixed effects in a **linear combination**: (effect*predictor) + (effect*predictor)…

Levels of Analysis in Two-Level Nested Data

- Between-Cluster (BC) Variation:
  - **Level-2 (L2)** = "**INTER**-cluster differences" = cluster characteristics
- Within-Cluster (WC) Variation:
  - **Level-1 (L1)** = "**INTRA**-cluster differences" = person characteristics
- **Any variable measured per person (at level 1)** likely has **both** level-2 between-cluster AND level-1 within-cluster variation!
  - BC = some clusters are higher/lower on average than other clusters
  - WC = some people are higher/lower than the rest of their cluster
- **So how do MLMs "handle" multiple levels of sampling?**
  - **TL;DR: by explicitly quantifying additional sources of (leftover) variability across *each* dimension of sampling in an outcome**

The Two Sides of *Any* Model

- Example GLM (p = person): $\boldsymbol{y_p} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{x1_p}) + \boldsymbol{\beta_2}(\boldsymbol{x2_p}) + \cdots + \boldsymbol{e_p}$

Our focus

- **Model for the Means:**
  - **Fixed Effects**, estimated **constant** contributions of predictor variables
  - What you are used to **caring about for testing hypotheses**
  - How the expected outcome for a given observation varies as a function of their values for the predictor variables (with the betas as estimated weights here)

- **Model for the Variance:**
  - **Residuals and Random Effects**, the "stochastic" or "error" part
  - What you are used to **making assumptions about** instead: $\boldsymbol{e_p} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma_e^2})$
  - How residuals are distributed and related across observations (persons, clusters, items, etc.) → these relationships are called "dependency" and ***this is the primary way that multilevel models differ from general linear models*** *(GLMs; "regression")*
    - Btw, random effect variances = latent variable variances (as in structural equation models)

"Empty" Single-Level vs Two-Level Models

- An empty (no-predictor) **single-level** model: $\boldsymbol{y_p = \beta_0 + e_p}$
  - All variability in the $\boldsymbol{y_i}$ outcome is captured by one "variance component" (or "pile" of variance in my vernacular; as in www.PilesOfVariance.com)
  - $\boldsymbol{e_p}$ = how far off a **person's** outcome is from the outcome **sample mean**, and the model estimates its variance across all people as $\boldsymbol{\sigma_e^2}$
  - Designed for single-level data because $\boldsymbol{e_p}$ is supposed to be independent across persons (persons as the p subscript here)

- An empty (no predictor) **two-level** model: $\boldsymbol{y_{pc} = \beta_0 + U_{0c} + e_{pc}}$
  - Variance in the $\boldsymbol{y_{pc}}$ outcome is now split into two "piles": $\boldsymbol{U_{0c} + e_{pc} \approx e_p}$
  - $\boldsymbol{U_{0c}}$ = how far off the **cluster's mean** is from the **sample mean of the cluster means**, with estimated variance across clusters as $\boldsymbol{\tau^2_{U_0}}$ (= "tau-squared sub U zero")
  - $\boldsymbol{e_{pc}}$ = how far off a person's outcome is from their cluster mean, with estimated variance across all people as $\boldsymbol{\sigma_e^2}$ (= "sigma-squared sub e"), *assumed equal across clusters*
  - **After controlling for clusters via $\boldsymbol{U_{0c}}$, $\boldsymbol{e_{pc}}$ *should* again be independent** (but maybe not!)

What can MLMs do for you?

1. **Model dependency across observations**
   - Longitudinal, clustered, and/or cross-classified data? No problem!
   - Tailor your model of sources of correlation to your data

2. **Include categorical or continuous predictors at any level**
   - Time-varying, person-level, cluster-level predictors for each variance
   - Explore reasons for dependency, don't just control for dependency

3. **Does not require same data structure for each unit**
   - Unbalanced clusters (or missing occasions)? No problem!

4. **You already know how (or you will soon)!**
   - What's an intercept? What's a slope? What's variance?

1. Model Dependency

- Sources of dependency depend on the sources of **variation** created by your sampling design: residuals for outcomes from the same unit are likely to be related, which violates the GLM "independence" assumption

- **"Levels" (for dependency)** ≈ "sources of variation due to sampling"
  - Sampling dimensions can be **nested**
    - e.g., time within person, person within cluster, school within district
  - If you can't figure out the direction of your nesting structure, odds are good you have a **crossed ("cross-classified") sampling design** instead
    - e.g., responses from persons crossed with items, ratings from raters crossed with targets
  - To have a "**level**", there must be **outcome variation due to sampling that is leftover** after including the model's fixed effects
    - e.g., treatment vs. control groups do not create another level of "clusters" (but it would if you had multiple treatment and multiple control groups)

Dependency arises from…

- Mean differences across sampling units (e.g., clusters)
  - Creates **constant** correlation across (exchangeable) persons from the same cluster
  - Will be represented by a **random intercept** in our MLMs (→ **homogeneous** variance)

- Cluster differences in slopes of person predictors
  - Creates **non-constant** correlation that depends on the person predictor values
  - Will be represented by **random slopes** in our MLMs (→ **heterogeneous** variance)

- Btw: MLMs can be extended to predict other heterogeneity of variance
  - These are called "location–scale mixed-effects models"
  - e.g., extent of within-team variance (= disagreement) differs across teams
  - e.g., extent of within-person fluctuation in mood differs across people

Why care about dependency?

- In other words, what happens if we have the wrong model for the variance (assume independent observations instead)?
- **Validity of the tests of the predictors** depends on having the "most right" model for the variance → *all cluster dependency is accounted for via random effects*
  - Estimates will usually be ok → come from model for the means, *so long as effects are disaggregated across levels of analysis!!!*
  - Standard errors (and thus *p*-values) will be compromised

- The sources of variation that exist in your outcome will dictate **what kinds of predictors** will be useful
  - Between-Cluster variation needs Between-Cluster predictors
  - Within-Cluster variation needs Within-Cluster predictors
  - Between-Item variation needs Between-Item predictors…

2 (of 3) Options for Cluster Dependency

Represent Cluster Differences via Fixed Effects

- Include (#clusters−1) binary predictors for cluster membership in the **model for the means** → **so clusters are then NOT a model "level"**
  - Main effects control for cluster mean differences only; interactions with person predictors are also needed to control for cluster slope differences
- Useful if #clusters < 10ish or you care about specific clusters, but then you cannot include cluster predictors → saturated mean differences

Represent Cluster Differences via Random Effects

- Include (at least) a random intercept variance across clusters in the **model for the variance** → **then clusters are a model "level"**
  - A random intercept controls for cluster mean differences only; a random slope variance is needed for cluster differences in person predictor slopes
- Better if #clusters > 10ish or you want to **predict** cluster differences

2. Include categorical or continuous predictors at any level of sampling (mainly applies to longitudinal)

- "ANOVA" test differences among discrete groups
- "Regression" tests slopes for continuous predictors
- What if a predictor is assessed repeatedly but can't be characterized by discrete "conditions"?
  - RM ANOVA or Regression won't work → you need MLM

- Some things don't change over time → time-invariant
- Some things do change over time → time-varying
- Some things are measured at higher levels
- Interactions are possible at same level or across levels

3. Does not require same longitudinal data structure per person (by accident or by design)

RM ANOVA: uses **multivariate** (wide) data structure and ordinary least squares (OLS) estimation:

ID	Sex	T1	T2	T3	T4
100	0	5	6	8	12
101	1	4	7	.	11

People missing any occasions are excluded (data from ID 101 here)

MLM: uses a **stacked** (long) data structure and likelihood-based estimation:

Only rows missing data are excluded
ID 100 uses 4 occasions
ID 101 uses 3 occasions

ID	Sex	Time	Y
100	0	1	5
100	0	2	6
100	0	3	8
100	0	4	12
101	1	1	4
101	1	2	7
101	1	3	.
101	1	4	11

Time can also be **unbalanced** across people such that each person can have their own measurement schedule: Time "0.9" "1.4" "3.5" "4.2"…

4. You already know how!

- If you can do GLM, you can do MLM (*and if you can do generalized linear models, you can do generalized multilevel models, too*)

- How do you interpret an estimate for…
  - the intercept?
  - the slope of a continuous variable?
  - the slope of a categorical variable?
  - a variance component ("pile of variance")?