Multilevel Modeling

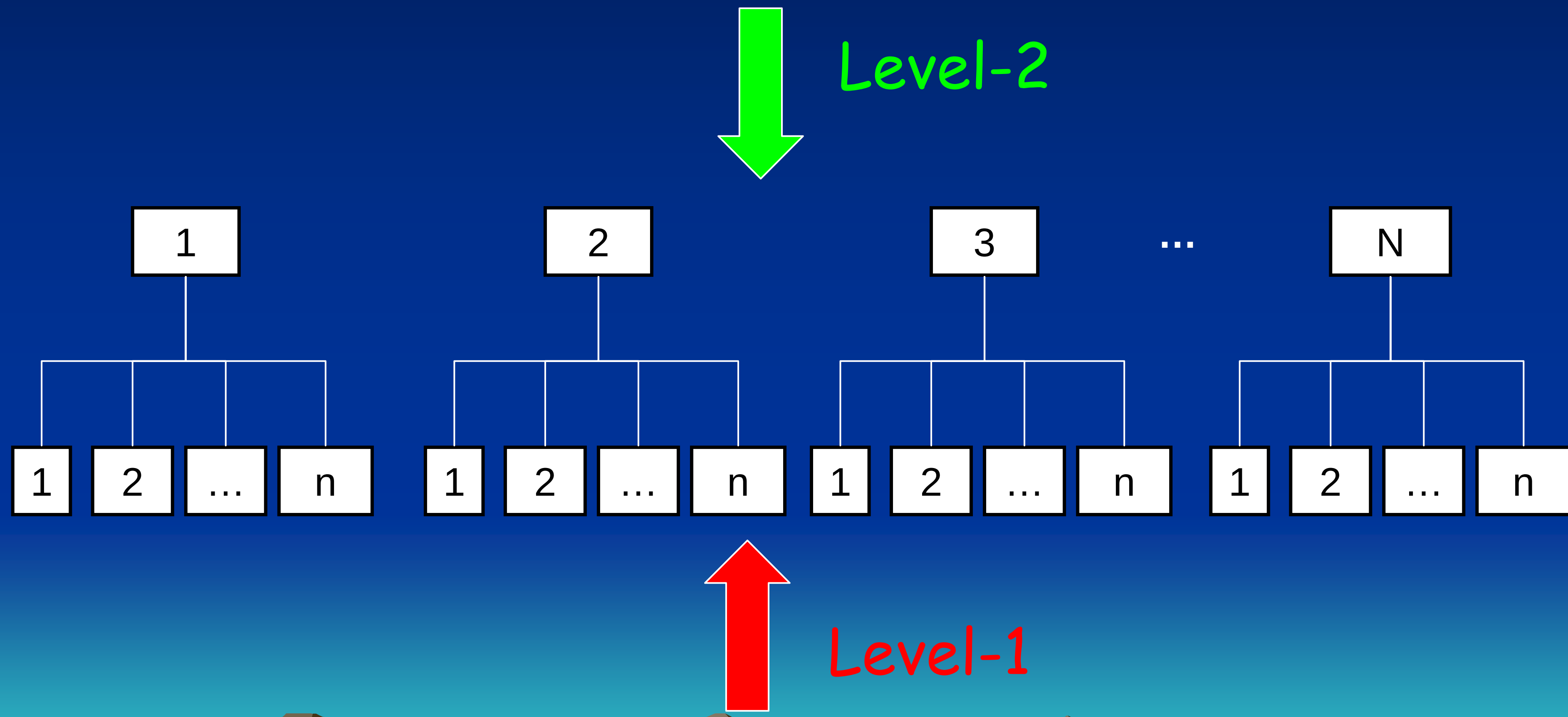
- Different titles for different disciplines
  - Sociology → multilevel linear modeling
  - Social Epidemiology → multilevel analysis (modeling)
  - Education → hierarchical linear modeling
  - Biometrics → mixed-effects or random-effects modeling
  - Statistics → covariance components
- Appropriate statistical model for
  - nested, grouped, or clustered data
  - longitudinal data

Multilevel Modeling

- Hierarchical linear modeling =
  - A statistical model that allows specifying and estimating relationships between variables
    - that have been observed at different levels of a nested data structure

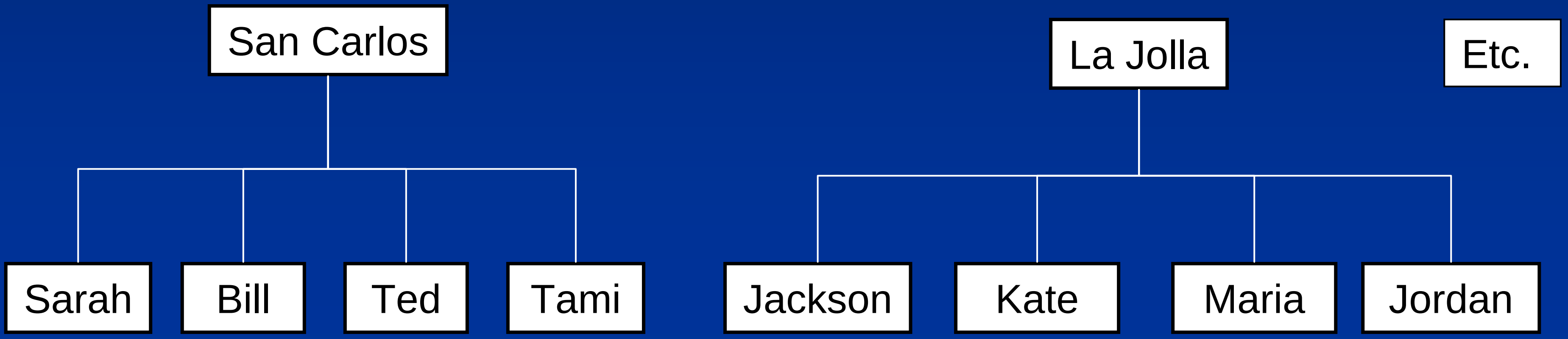
Nested data structures

- Nested data structures are everywhere
  - 2-level data structure



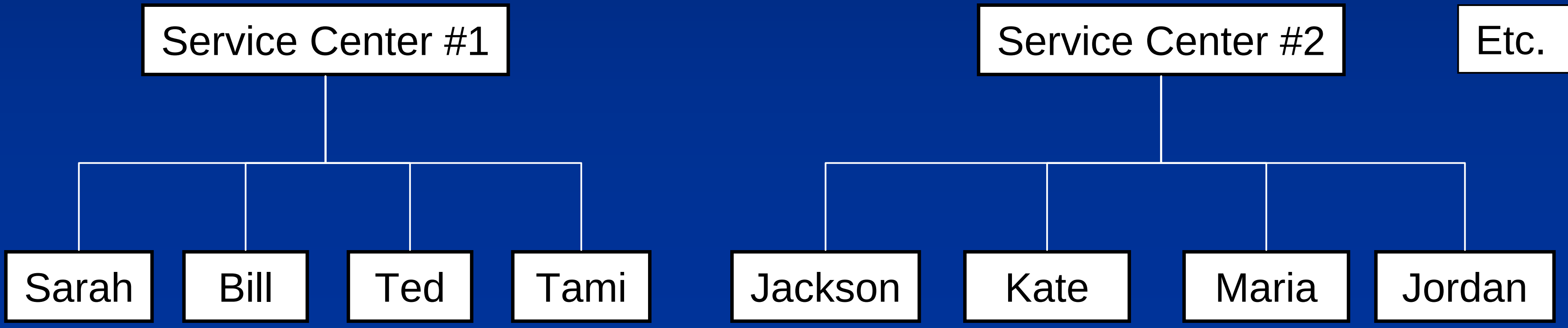
Nested Data Structures

- Grouped data
  - Individuals nested within neighborhoods (communities)



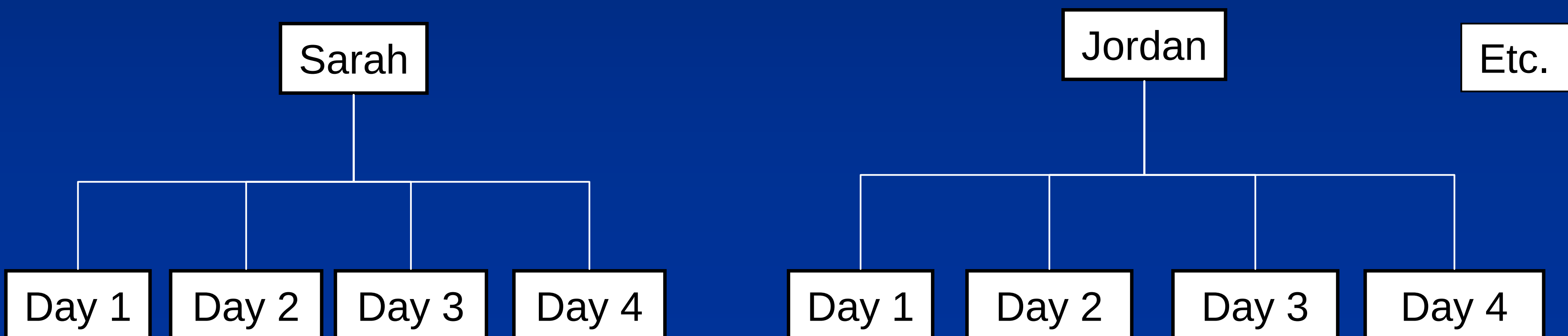
Nested Data Structures

- Grouped data
  - Patients nested within mental/physical health service centers



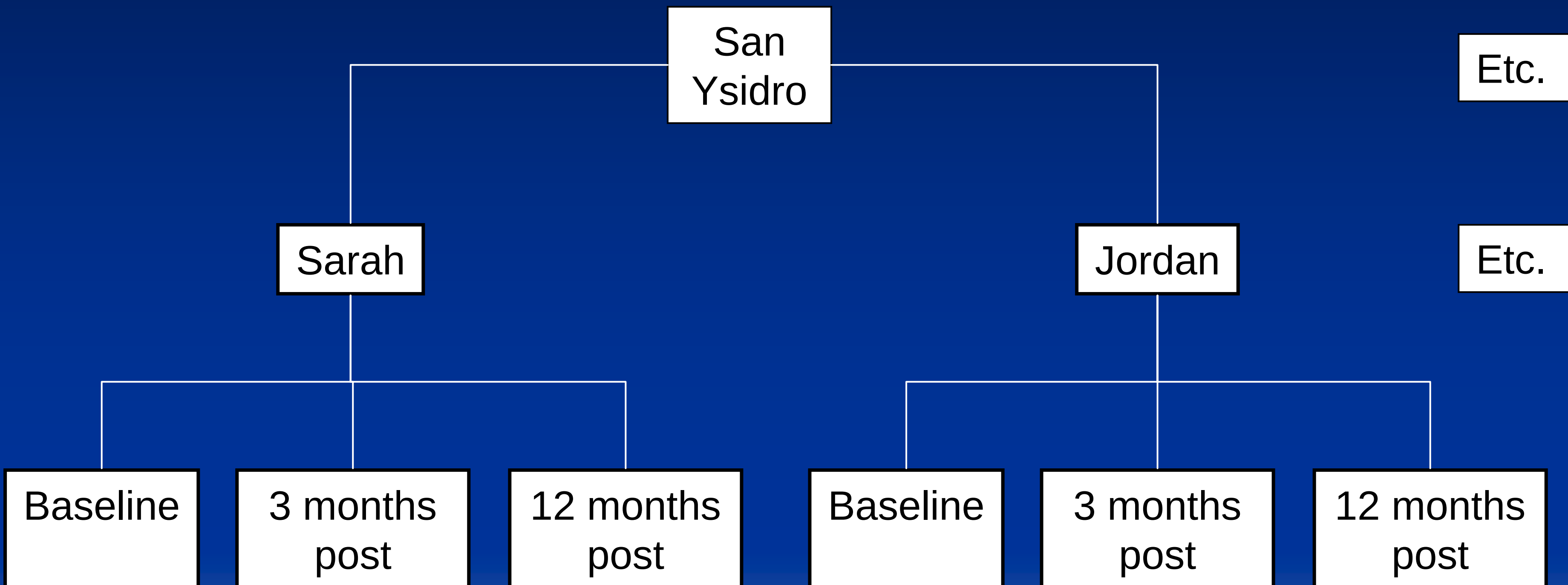
Nested Data Structures

- Longitudinal data
  - Time periods (or repeated observations) nested within individuals (2-level structure)



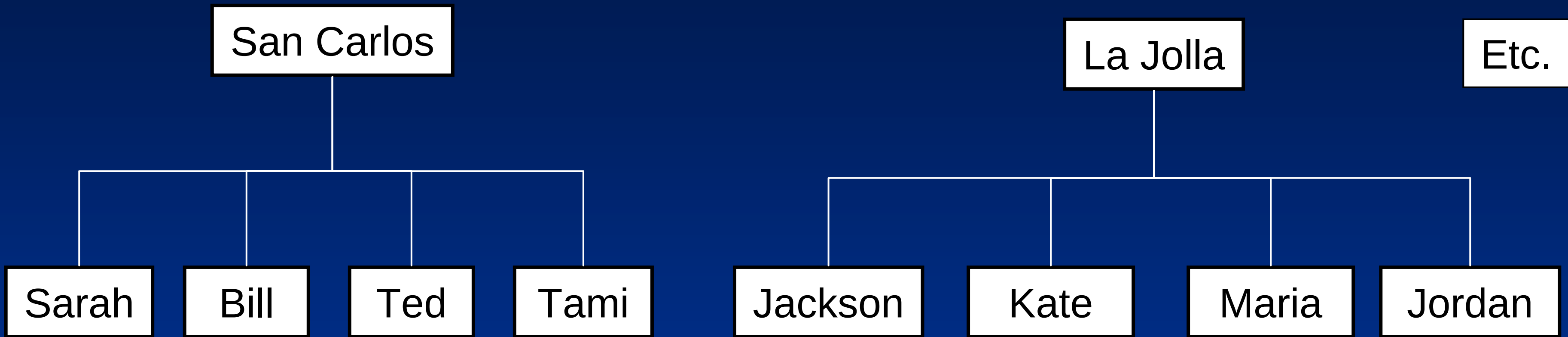
Nested Data Structures

- 3-level nested data structure



- repeated observations (level-1) nested within individuals (level-2) nested within communities (level-3)

Problems with Nested Data Structures



- Suppose we have an outcome at level-1
  - e.g., participants' physical activity levels
- We want to predict this outcome with
  - participants' Extraversion (level-1 variable)
  - community-level cohesion (level-2 variable)

Problems with Nested Data Structures

- Many questions arise
  - How do we do this?
  - What is the proper sample size?
  - What if the effect of Extraversion on physical activity is NOT the same across communities?
- Before I answer these questions let's discuss what traditionally has been done

Traditional Approaches and their Problems

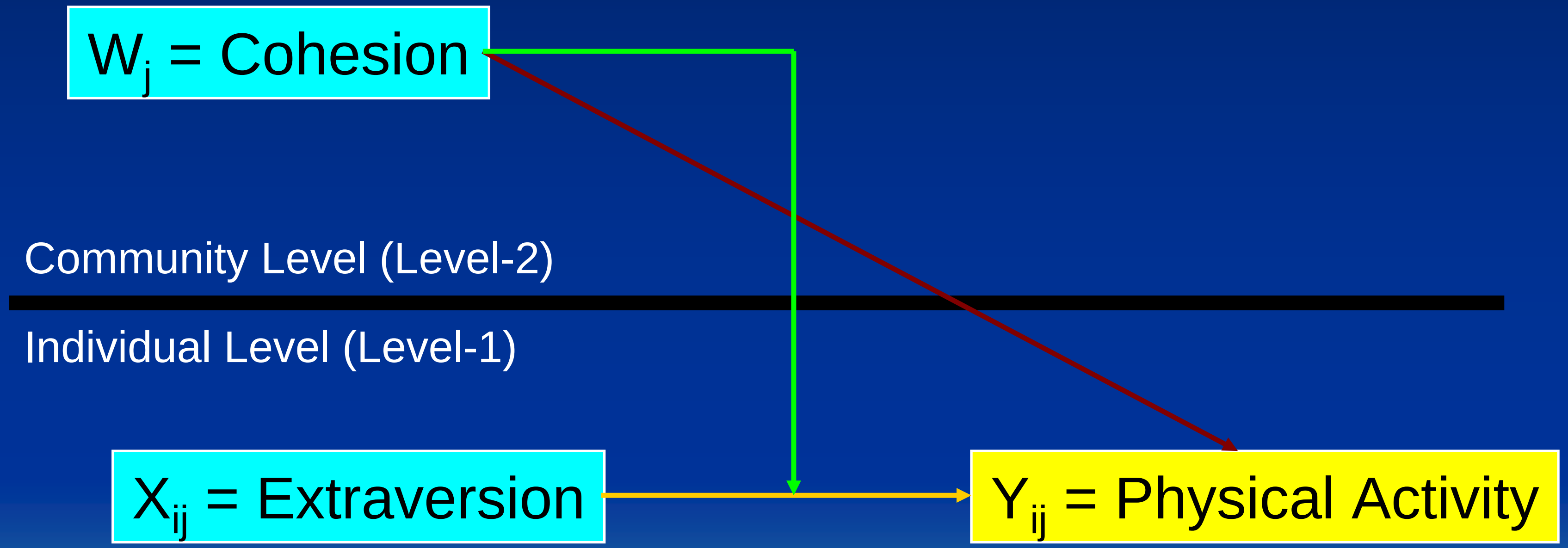
- Disaggregation method
  - analyze individual-level scores ignoring the hierarchical data structure
  - e.g., assigning every individual WITHIN the same community the same group cohesion value
  - run a basic OLS regression (or ANOVA)
  - primary problem
    - participants treated as independent entities
      - violates independence assumption, which ↓ standard errors
      - participants WITHIN a community are MORE alike

Traditional Approaches and their Problems

- Aggregation method
  - averaging individual-level scores for each level-2 unit
    - e.g., averaging both the physical activity and the Extraversion scores for individuals WITHIN a community to create community-level variables ONLY
  - run a basic OLS regression (or ANOVA)
  - primary problem
    - greatly reduces your sample size, and thus power
    - throws away within-in group information

Graphical Representation of a Simple Hierarchical Linear Model

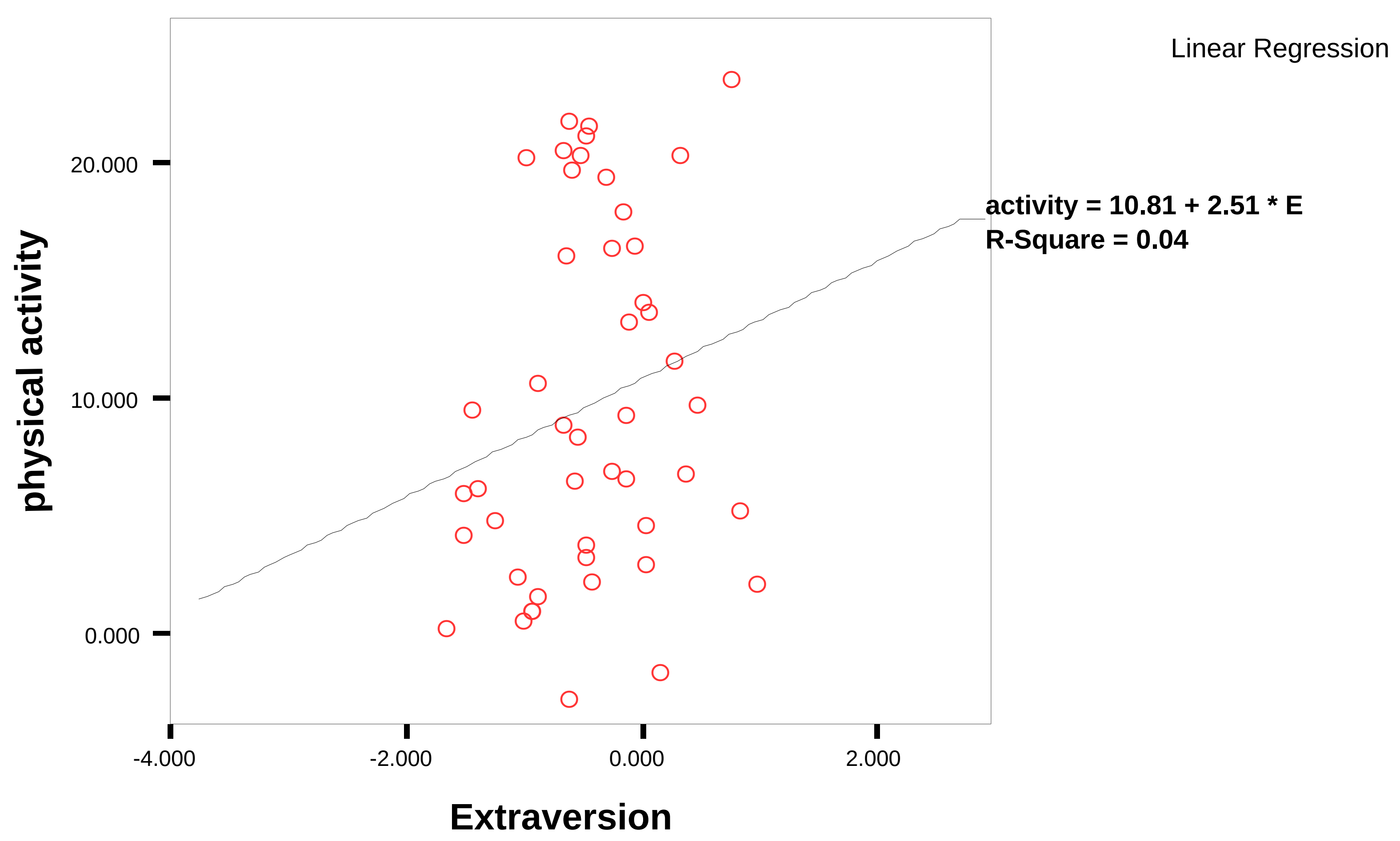
- We want a technique that allows for variables and variation at all levels of the data structure



The logic of HLM

- Continuing with our example
  - let's examine the basic regression equation in a single community
    - $Y' = \beta_0 + \beta_1 X_i + r_i$
    - Physical Activity' = $\beta_0 + \beta_1(\text{Extraversion})_i + r_i$
      - assume that we have grand-mean centered Extraversion
    - Let's graph this regression equation on the next slide

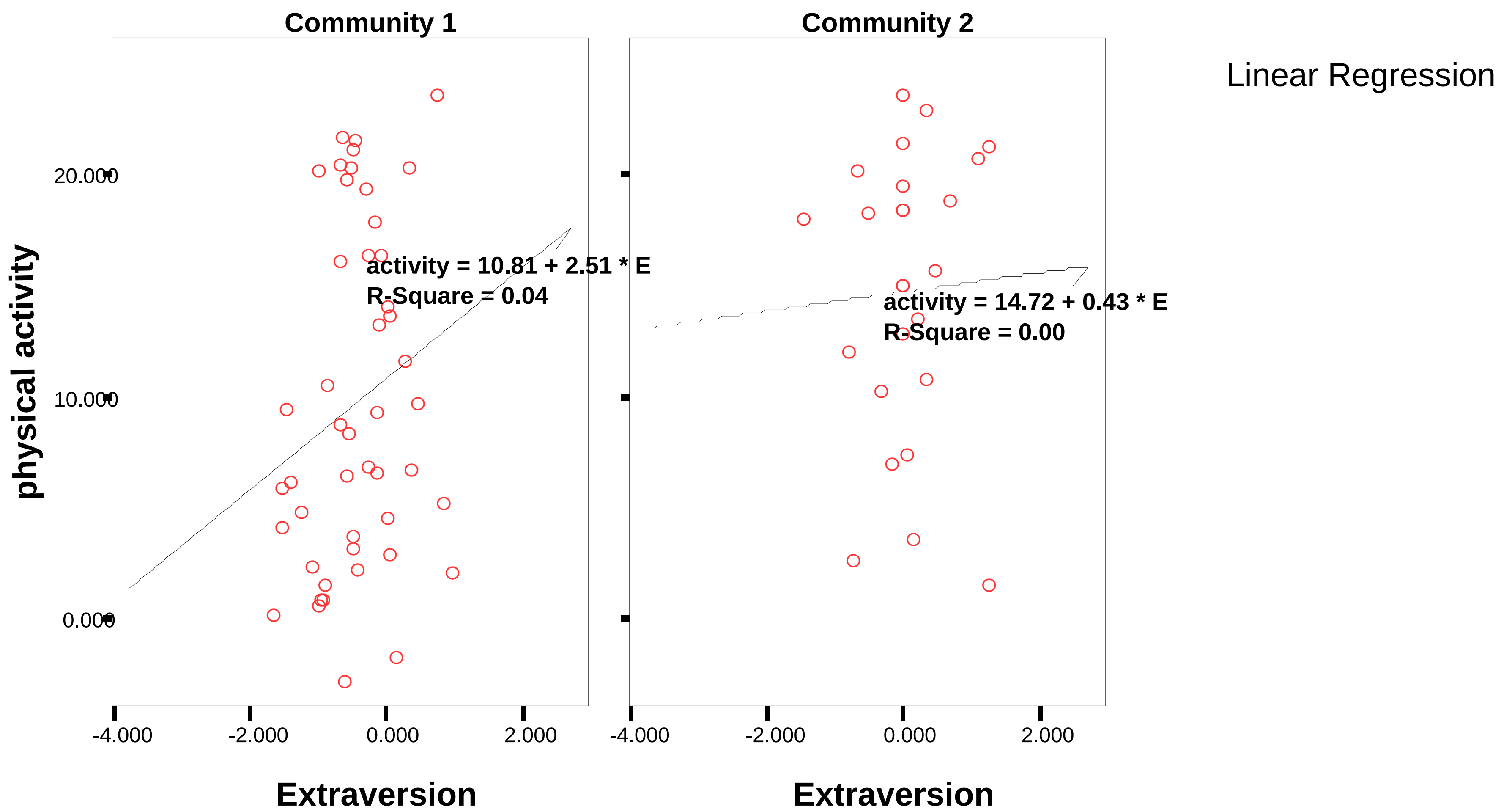
Regression in a single community



Extending this to 2 Communities

- Community 1
  - Physical Activity' = $\beta_{01} + \beta_{11}(\text{Extraversion})_i + r_i$
- Community 2
  - Physical Activity' = $\beta_{02} + \beta_{12}(\text{Extraversion})_i + r_i$
- So each community has its own intercept and slope
  - this in effect serves as further "data"
    - more on this to come
- Let's graph these equations on the next slide

Regression Line in 2 Communities



Multiple Level-2 Units

- What if we consider multiple communities?
  - Physical Activity$'_{ij} = \beta_{0j} + \beta_{1j}$(Extraversion)$_{ij} + r_{ij}$
    - intercept and slope are referenced by j, which means each community has their own values
    - two important pieces of information when we summarize values across level-2 units
      - mean intercept value and the Physical Activity-Extraversion slope
      - how different each level-2 unit is from these respective means

Terminology: random and fixed

- Fixed effects are variable coefficients that are constant across level-2 units
  - mean intercept and slope across level-2 units
    - we have statistical tests for these
- Random effects are coefficients that can vary across level-2 units
  - error terms for level-1 and level-2 equations
    - we have statistical tests for these

Hierarchical Linear Regression Model

- Each level of the nested data structure is represented by its own submodel
- Assume individuals (level-1) are nested within communities (level-2)
  - level-1 DV = individuals' physical health (Y_{ij})
  - level-1 IV = individuals' Extraversion (X_{ij})
  - level-2 IV = community group cohesion (W_j)
  - Level-1 Equation:
    - Physical Activity$_{ij}$= β_{0j}+ β_{1j}(Extraversion$_{ij}$)+ r_{ij}
  - Level-2 Equation***s***:
    - intercept (β_{0j}) = γ_{00}+ γ_{01}(Cohesion$_j$) + u_{0j}
    - slope (β_{1j}) = γ_{10}+ γ_{11}(Cohesion$_j$) + u_{1j}

Hierarchical Linear Regression Model

- we can create a single equation representing the regression model by
  - substituting the level-2 equations into the level-1 equation and using a little algebra
  - $PA_{ij} = \gamma_{00} + \gamma_{10}(E_{ij}) + \gamma_{01}(Cohesion_j) + \gamma_{11}X_{ij}Z_j + \ldots$
    - fixed effects part of the equation
  - $PA_{ij} = \ldots r_{ij} + u_{0j} + u_{1j}X_{ij}$
    - random effects part of the equation

- Let's break this down using different types of models and the questions that they ask and answer

The intercept-only (unconditional) model

- Primary questions this model answers
  - What is the mean outcome score aggregated across level-2 units?
  - How much variability is there for the outcome at each level of the data structure?
- Equations
  - Level-1 equation: $PA_{ij} = \beta_{0j} + r_{ij}$
    - β_{0j} = mean PA score for each community
    - r_{ij} = difference between each individual's PA and their community's average PA
      - we are interested in the variance of this residual term

The intercept-only (unconditional) model

- level-2 equation: $\beta_{0j} = \gamma_{00} + u_{0j}$
  - β_{0j} = PA score for each community
  - γ_{00} = mean PA score across communities
    - referred to as the grand mean
  - u_{0j} = difference between each community's PA and the grand mean for PA
    - the variance of this residual term is of primary interest

The intercept-only (unconditional) model

- Intraclass correlation coefficient (ICC or ρ)
  - indicates the amount of variance in the outcome that is a function of level-2 units
    - i.e., what proportion of variance in the outcome is explained by *differences* in the level-2 units

$$\text{ICC} = \frac{\text{Variance of the difference between each level-2 mean and the grand mean}}{\text{Total Variance}}$$

  - e.g., ICC = .27 means that 27% of the PA differences are a function of the community that one lives in
    - suggests that individuals within communities are NOT independent

The means-as-outcomes model

- Primary questions this model answers
  - Does the mean outcome score differ as a function of level-2 predictor variables?
  - Do level-2 units still differ with respect to the outcome after accounting for predictor variables already in the model?
- Equations
  - Level-1 equation: $PA_{ij} = \beta_{0j} + r_{ij}$

The means-as-outcomes model

- Level-2 equation: $\beta_{0j} = \gamma_{00} + \gamma_{01}(Cohesion_j) + u_{0j}$
  - β_{0j} = mean PA score for EACH community
  - γ_{00} = mean PA for those in communities that have "average" cohesion
  - γ_{01} = association between a community's PA and their community-level cohesion
  - variance component
    - the variance of the residual (u_{0j}) is the difference between each community's PA and grand mean PA, after accounting for community cohesion
    - indicates whether communities still differ

The random-coefficients regression model

- Primary questions this model answers
  - What is the mean intercept and slope across level-2 units?
  - How much do the intercept and slope differ among level-2 units?
- Equations
  - Level-1 equation: $PA_{ij} = \beta_{0j} + \beta_{1j}(E_{ij}) + r_{ij}$
    - β_{0j} = PA score for individuals who are "average" on Extraversion
    - β_{1j} = association between individual-level Extraversion and PA
    - variance r_{ij} = individual PA scores still differ?

The random-coefficients regression model

- Level-2 equations
  - $\beta_{0j} = \gamma_{00} + u_{0j}$
    - γ_{00} = grand mean PA score controlling for individual level Extraversion
    - variance u_{0j} = community PA scores still differ?
  - $\beta_{1j} = \gamma_{10} + u_{1j}$
    - γ_{10} = association between Extraversion and PA aggregated across communities
    - variance u_{1j} = association different across communities?

Intercepts- and slopes-as-outcomes model

- 2 primary questions to answer
  - Why some level-2 units have higher/lower outcome scores than others (intercept)?
  - Why some level-2 units have a stronger/weaker level-1 relations than other level-2 units?
- Level-1 equation: $PA_{ij} = \beta_{0j} + \beta_{1j}(E_{ij}) + r_{ij}$

Intercepts- and slopes-as-outcomes model

- Level-2 equations
  - $\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Cohesion}_j) + u_{0j}$
    - γ_{00} = mean community PA for communities who have average cohesion
      - assume we have grand-mean centered the cohesion predictor
    - γ_{01} = association between community-level PA and community-level cohesion
    - variance u_{0j} = community PA scores still differ?

Intercepts- and slopes-as outcomes model

- $\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{Cohesion}_j) + u_{1j}$
  - γ_{10} = association between individual-level Extraversion and PA aggregated across communities
  - γ_{11} = association between individual-level Extraversion and PA is predicted by community-level cohesion
    - cross-level interaction
  - variance u_{1j} = association different across communities?
- can also evaluate the covariance (correlation) between the intercepts and slopes

Practical Issues

- Type of outcome variables
  - multilevel analogs for each
    - binary → logistic regression
    - unordered categorical → multinomial logistic regression
    - ordered categorical → ordinal regression
    - counts → Poisson regression
- Regression coefficients
  - standardized vs. unstandardized
- Deviance statistic

Practical Issues continued

- Centering predictor variables (see Raudenbush & Bryk [2002], pp. 31-35 & 134-149 for a full explanation)
  - uncentered
  - grand-mean
  - group-mean
- Computer programs for analysis
  - HLM, MLwiN, MPlus, SPSS (Mixed Models), SAS (Proc Mixed), MIXOR, Stata
- Power analysis
  - "30/30" and "50/20" rule
  - computer programs include RMASS2, Optimal Design, PINT, MPlus

Primary Book References

- Heck, R.H., & Thomas, S.L. (2000). *An introduction to multilevel modeling techniques*. Mahwah, NJ: Erlbaum
- Hox, J.J. (2002). *Multilevel analysis*. Mahwah, NJ: Erlbaum.
- Kreft, I., & DeLeeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage
- Raudenbush, S.W., & Bryk, A.S. (2001). *Hierarchical linear models*. 2nd edition. Newbury Park, CA: Sage.
- Singer, J.D., & Willet, J.B. (2003). *Applied longitudinal data analysis*. Oxford, England, Oxford University Press.
- Snijders, T.A.B., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced modeling*. London, England: Sage.

Useful Websites and Discussion Lists

- Center for Multilevel Modeling
  - http://www.mlwin.com
- UCLA Statistical Computing Center
  - http://statcomp.ats.ucla.edu/mlm/
- Multilevel Modeling Discussion List
  - http://www.jiscmail.ac.uk/lists/multilevel.html