

Mixed Models

Workshop starts at 12:02

No coding required

We'll make demo script available afterwards

Follow @Northwestern_IT on Instagram for the latest workshop and bootcamp updates.



Introduction to Mixed Models for Correlated Data

Instructor: David Nichols

TA: Emilio Lehoucq

This workshop is brought to you by:

Northwestern IT Research Computing and Data Services

Need help?

- AI, Machine Learning, Data Science
- Statistics
- Visualization
- Collecting web data (scraping, APIs), text analysis, extracting information from text
- Cleaning, transforming, reformatting, and wrangling data
- Automating repetitive research tasks
- Research reproducibility and replicability
- Programming, computing, data management, etc.
- R, Python, SQL, MATLAB, Stata, SPSS, SAS, etc.

Request a **FREE** consultation at bit.ly/rcdsconsult.

What's next?

- October the 29th: Next Steps in R: Writing Your Own Functions
- November the 4th: Next Steps in Python: Virtual Environments
- November the 12th: AI for Research: Choosing a LLM for Your Project
- November the 19th: Introduction to SQL (in-person)
- November the 20th: Intermediate SQL (in-person)
- December 16 – 19: Bootcamp: R Fundamentals
- December 16 – 19: Bootcamp: Python Fundamentals

Logistics

- **Ask Questions** [in the zoom chat].
 - If you know the answer, feel free to respond (we may politely clarify if needed).
 - Emilio will monitor and we will address questions after the presentation.
- **If my internet goes out.**
 - Take a 5 minute break, and we will meet back in the same zoom room.



Introduction and Goals

Introduction to Mixed Models

Short Description: An introduction to the use of mixed models for handling correlated data in regression modeling

Longer Description: Statistical models incorporating random effects, known as mixed models, provide options for modelling correlated data that violate the assumptions of simpler models. This workshop will introduce several types of mixed models that can handle many common data types, such as repeated measures, longitudinal data, and clustered data. This is primarily a theoretical workshop, not a coding one, but tips on how to implement mixed models in statistical software will be included.

Mixed Models

Prerequisites: Knowledge of basic statistical methods and some experience with regression models

Everyone interested is welcome, but if you don't have some experience with at least linear regression models, you probably won't have an easy time following the presentation.

There will also be some use of matrix notation.

Mixed Models

Goals of the workshop:

- Explain why it's so important to address lack of independence among observations
- Show how mixed models address these issues
- Illustrate two of the most common linear mixed models



The Basic Linear Model

The univariate fixed effects linear model

The univariate fixed effects linear model is

$$y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \varepsilon_i$$

- y_i is the observed dependent variable value for the i^{th} case
- β_0 is a fixed but unknown intercept or constant
- β_k is a fixed but unknown parameter for the k^{th} predictor variable
- X_{ik} is the known fixed value of the i^{th} case for the k^{th} predictor
- ε_i is a random error for the i^{th} case

The univariate fixed effects linear model

- The X_{ij} may be instances of “continuous” variables, which are also known as quantitative, interval, measured, or scale variables, or covariates
- They may also be indicators indexing the levels of categorical variables or *factors* representing separate groups of cases, or other encodings representing comparisons among groups
- They may also be variables representing products of elemental variables (interactions between variables) or powers of a single variable (polynomial terms)

The standard normal theory linear model

So far nothing has been said about the assumptions related to the random errors ε_i

For the standard model, estimated using ordinary least squares (OLS), to apply t and F tests, these assumptions can be stated as

$$\varepsilon_i \sim i.i.d. \mathcal{N}(0, \sigma^2)$$

This means the errors ε_{ij} are independently and identically distributed (“*i. i. d.*”) according to a normal distribution with mean or expectation 0 and common variance σ^2

Standard linear model assumptions

These assumptions are often stated as:

1. Independence
2. Homogeneity or equality of variance
3. Normality

What are the relative importances of these assumptions for the use of standard t and F tests and associated confidence intervals?

Relative importance of assumptions

There is generally a clear hierarchy of importance:

1. Independence
2. Homogeneity or equality of variance
3. Normality



Correlated Data and Its Effects

Examples of correlated data

- Responses to survey questions from members of the same family
- Test scores from students in the same classroom
- Measurements taken from related animals or members of a pack or group
- Plant measurements taken from plants sharing a common ecological area
- Any measurements or responses where the same thing is being observed multiple times (often referred to as repeated measures or longitudinal data, depending on spacing of observations)

Common sources of correlated errors

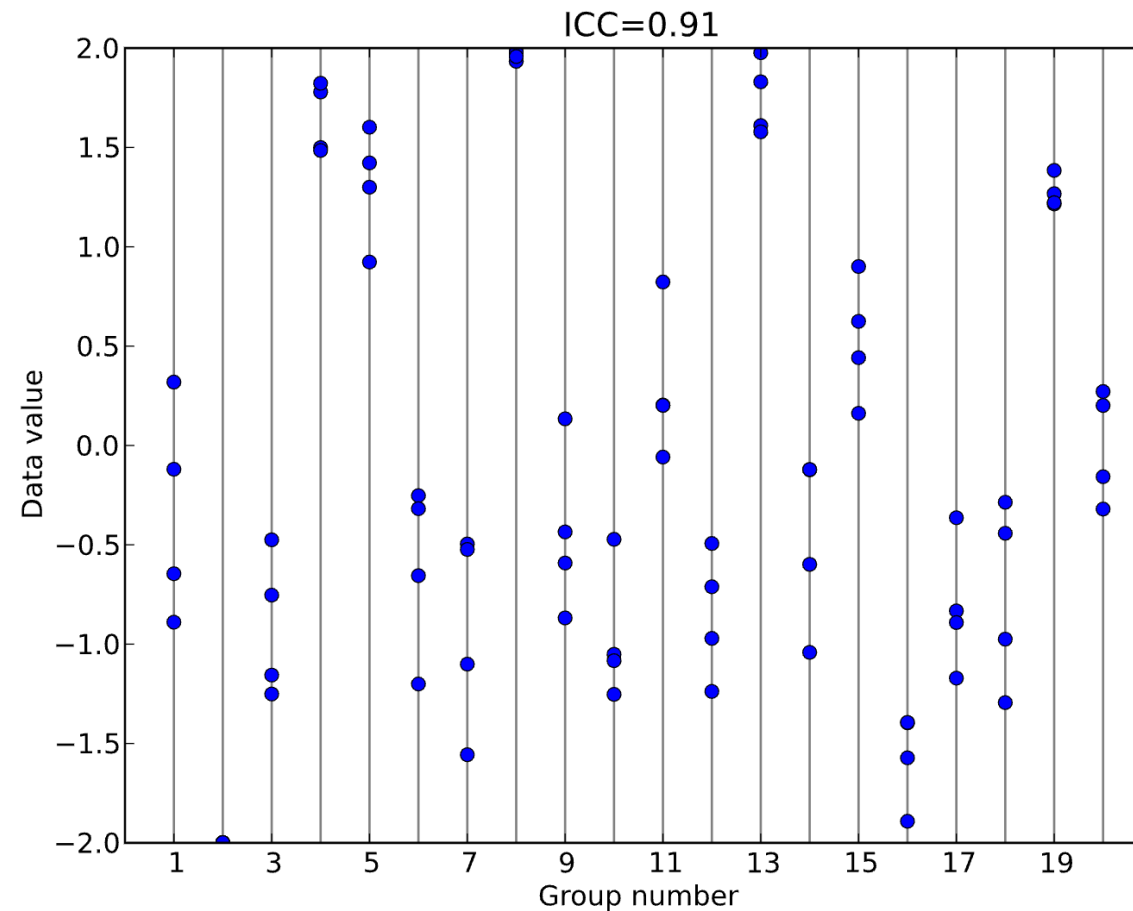
The most common reasons for correlated errors are:

- Units being sampled in groups rather than independently
- Units sampled independently being exposed to common conditions not included in the assumed model
- Repeated measurements taken on the same cases or units

Correlations are usually positive

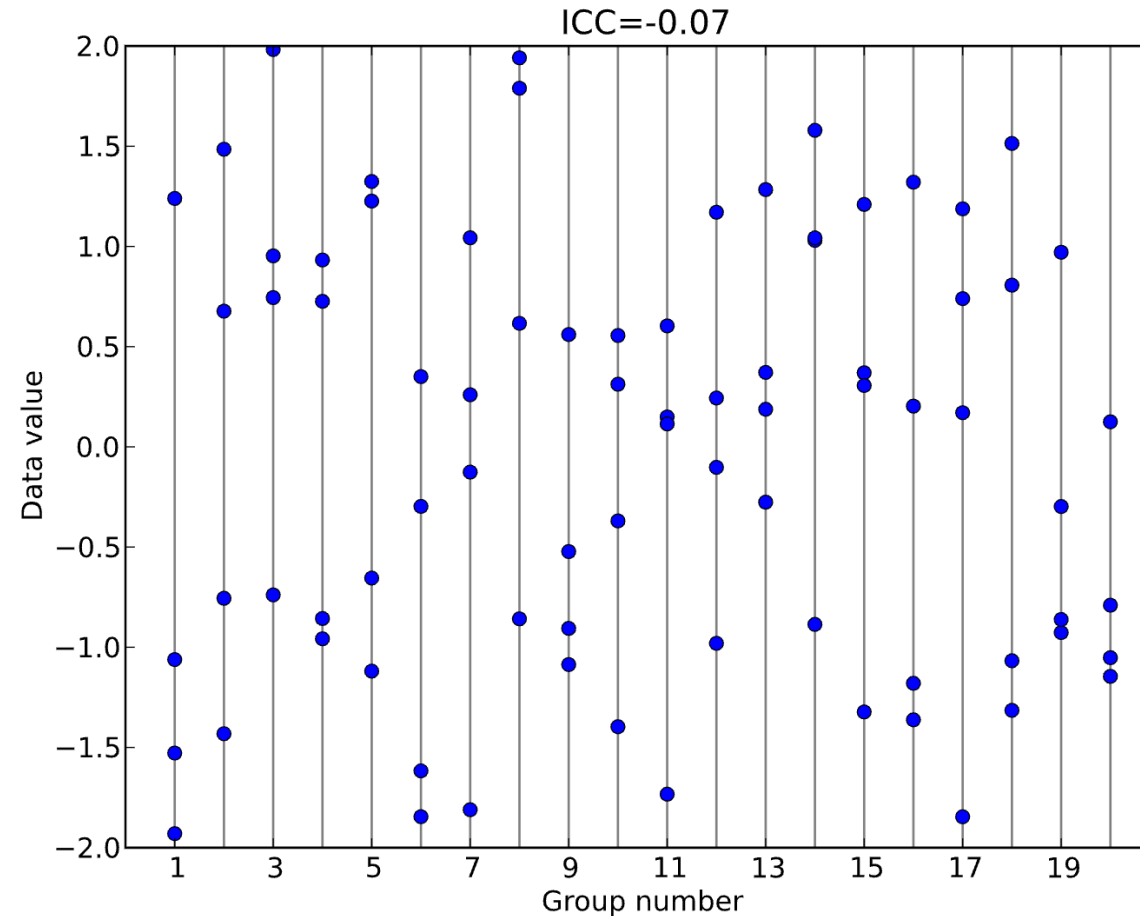
- In the examples on the earlier slide, observations on units within the same grouping, area, or repeated on the same individual are likely to be positively correlated, or more alike, than observations drawn at random from across the groupings, areas, or distinct individuals.
- The extent of relatedness is measured by an *intraclass correlation (ICC)*, usually denoted by ρ . Intraclass correlations index the extent to which observations randomly drawn from the same group are more (or rarely, less) alike than observations randomly drawn from different groups.

Example data with a high ICC



By Skbkekas - Own work, CC BY-SA 3.0,
<https://commons.wikimedia.org/w/index.php?curid=6982852>

Example data with a small negative ICC



By Skbkekas - Own work, CC BY-SA 3.0,
<https://commons.wikimedia.org/w/index.php?curid=6982834>

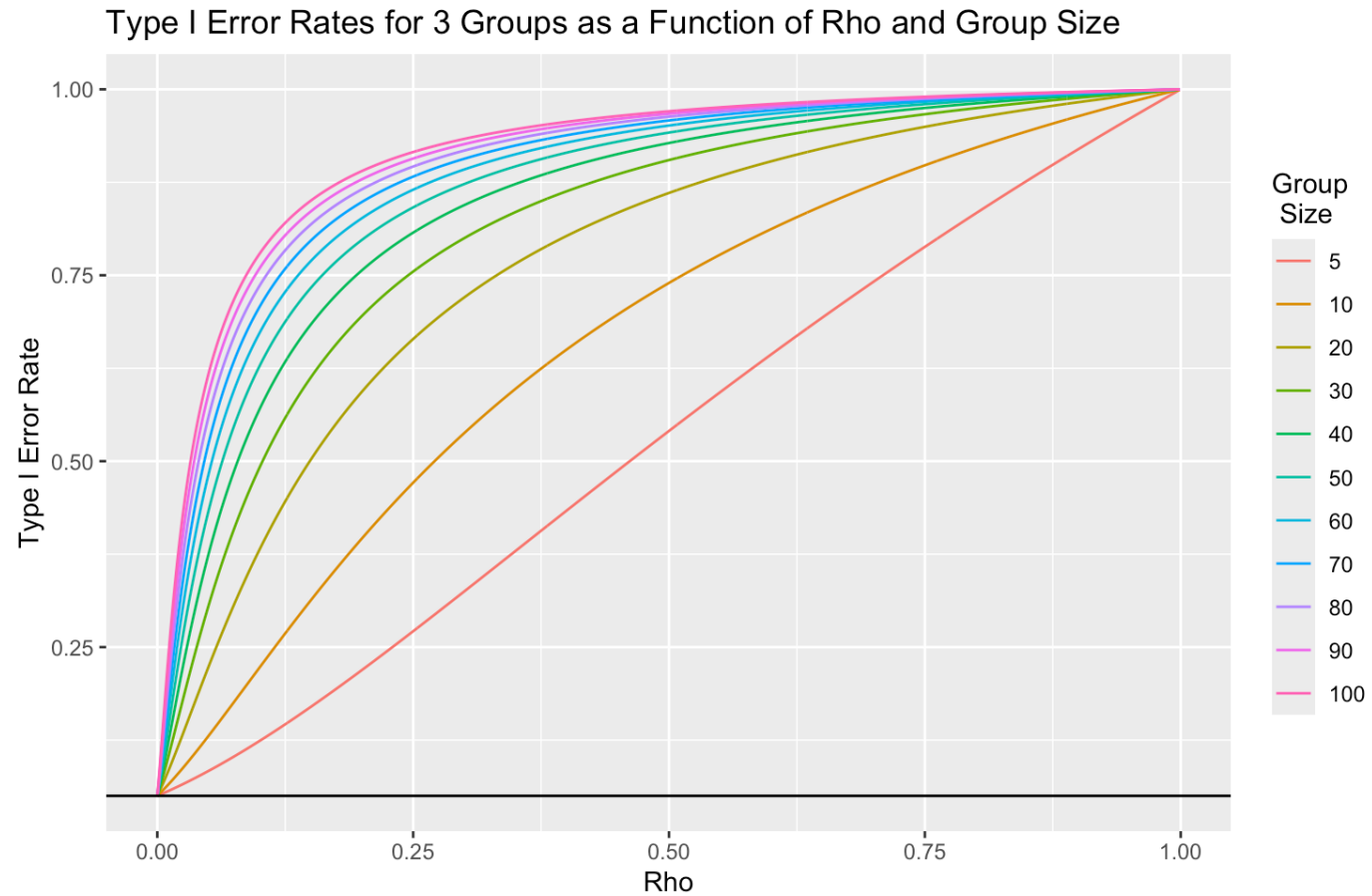
Illustrating the effects of positively correlated errors

- The following three slides graph Type I error rates on the vertical or Y axis against intraclass correlation values (ρ) on the horizontal or X axis, for situations with 2, 3, and 5 groups of equal sample sizes
- In each case, data are drawn from populations with equal population variances, normally distributed errors, and 0 population differences between groups
- Thus all assumptions other than independence are strictly met

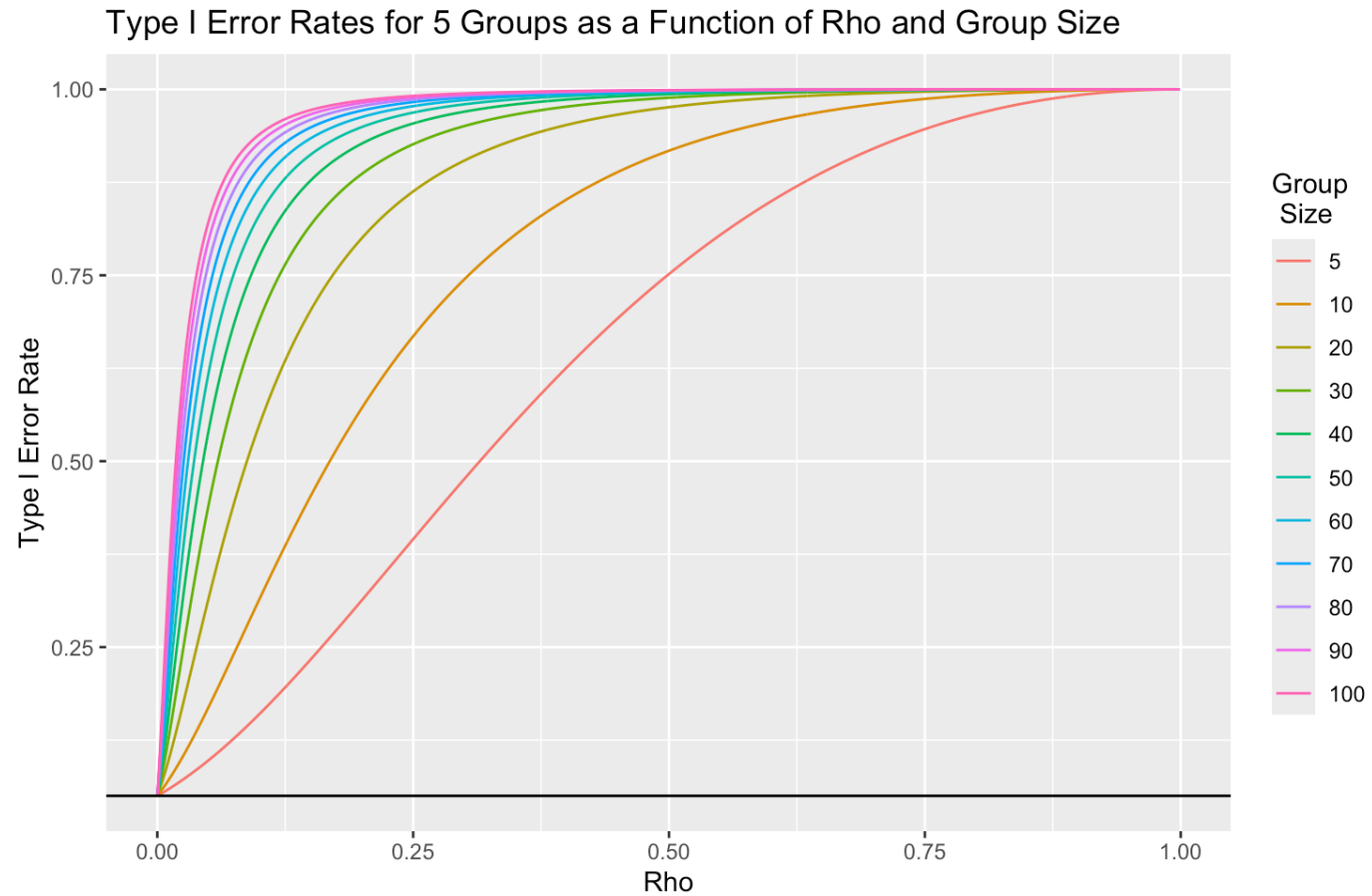
Type I error rates with 2 groups



Type I error rates with 3 groups



Type I error rates with 5 groups



Implications of positive intraclass correlation

- Effects on Type I error can be dramatic
- For a given number of groups and observations per group, as ρ increases, the Type I error rate increases and can approach 1
- For a given ρ , increasing the sample size either by adding more groups of the same size or increasing the sample size within groups results in increases in the Type I error rate
- Effects on confidence interval coverage levels are equally bad: as the true ρ level increases the percentage of $100 * (1 - \alpha)\%$ confidence intervals that cover the true 0 value decreases and can approach 0%

Implications of positive ICC

- Effects on Type I error can be dramatic and are typically much more severe than those associated with inequality of variance or non-normality
- Unlike heterogeneity of variance or non-normality, *increasing sample sizes does not help – it makes things worse!*
- The central limit theorem doesn't help here because it assumes independence of observations



Handling Correlated Errors with Mixed Models by Adding Random Effects

Data structure for mixed models

- An ID variable is required to link cases for a subject
- Repeated measurements are held in the same variable over different cases with the same ID
- An index variable is used to identify within-subject structured observations (e.g., repeated measures)
- This structure is referred to as the univariate or mixed models approach setup, or as long or narrow data

ID	group	time	y
1	1	1	3.0
1	1	2	3.6
1	1	3	4.4
1	1	4	5.1
2	1	1	2.5
2	1	2	2.8
2	1	3	3.2
2	1	4	3.9
3	1	1	3.5
3	1	2	3.9
3	1	3	4.2
3	1	4	5.1

The univariate fixed effects linear model again (just for easy reference to next slide)

The univariate fixed effects linear model is

$$y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik} + \varepsilon_i$$

- y_i is the observed dependent variable value for the i^{th} case
- β_0 is a fixed but unknown intercept or constant
- β_k is a fixed but unknown parameter for the k^{th} predictor variable
- X_{ik} is the known fixed value of the i^{th} case for the k^{th} predictor
- ε_i is a random error for the i^{th} case

Fixed effects univariate linear model: Matrix representation

The matrix representation of the model presents a more compact form:

$$y = X\beta + \varepsilon$$

- y is a column vector of observed values with N rows
- β is a fixed but unknown column vector of parameters for the $k + 1$ parameters in the model
- X is an $N \times (k + 1)$ matrix of known fixed predictor variable values
- ε is a column vector of random errors with N rows

Fixed effects linear model: Matrix representation

The assumptions about the random errors

$$\varepsilon_i \sim i.i.d. \mathcal{N}(0, \sigma^2)$$

are

$$E(\varepsilon) = 0$$

$$Var(\varepsilon) = R = \sigma^2 I$$

Linear mixed models: Matrix representation

The linear mixed model equation is

$$y = X\beta + Zb + \varepsilon$$

- b is a column vector of length qn
- Z is an $N \times qn$ design matrix for the random effects
- q is the number of random effects, n is the number of *subjects*, and N is the total number of observations over all subjects

Linear mixed models assumptions

The random effects b are assumed to distributed as

$$b \sim \mathcal{N}(0, G)$$

the errors ε are assumed to be distributed as

$$\varepsilon \sim \mathcal{N}(0, R)$$

and b and ε are assumed to be uncorrelated

Linear mixed models assumptions

$$E \begin{bmatrix} b \\ \varepsilon \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$Var \begin{bmatrix} b \\ \varepsilon \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}$$

Linear mixed models: Drilling down

- The model presented in the previous slides is a very general one that encompasses a broad range of possible models that allow for both correlations among observations on the same subjects and also unequal variances across groups of subjects and/or times
- Not all software procedures or packages for mixed models support this level of generality (those that do include SAS and SPSS)
- For this introduction, we will show examples of two common, simple mixed models, the random intercept model for clustered data, and the random slope and intercept model for repeated measurements or longitudinal data



The Random Intercept Model for Clustered Data

Random intercept model

- Although this model can also be used for other data, including repeated measures and longitudinal data, it is most commonly used when data have a clustered structure where observations within a given cluster (or subject) are correlated, but there is no further within-subject structure
- *The addition of a random intercept term introduces a constant intraclass correlation among observations within subjects*
- Observations for different subjects remain independent (the R matrix is modeled as a diagonal matrix with the residual variance in each diagonal element, just as with the standard regression model)

Random intercept model

The Z (or random effects design) matrix has a set of 0-1 indicator or dummy variables, one for each of the n subjects, so each observation within a given subject i shares an additional, random intercept b_{0i} :

$$y_{ij} = \beta_0 + \beta_1 X_{ij1} + \dots + \beta_k X_{ijk} + b_{0i} + \varepsilon_{ij}$$

- y_{ij} is the observed dependent variable value for the j^{th} observation within the i^{th} subject
- X_{ijk} is the known fixed value of the k^{th} predictor for the j^{th} observation within the i^{th} subject

Random intercept model

Another way to express the equation is to group the fixed and random intercepts together as

$$y_{ij} = (\beta_0 + b_{0i}) + \beta_1 X_{ij1} + \dots + \beta_k X_{ijk} + \varepsilon_{ij}$$

- β_0 represents the population-averaged (or “marginal”) mean
- $\beta_0 + b_{0i}$ represents the mean for subject i conditional on the random effects
- $E[b_{0i}] = 0$, so b_{0i} represents the deviation of subject i from β_0

Random intercept model example

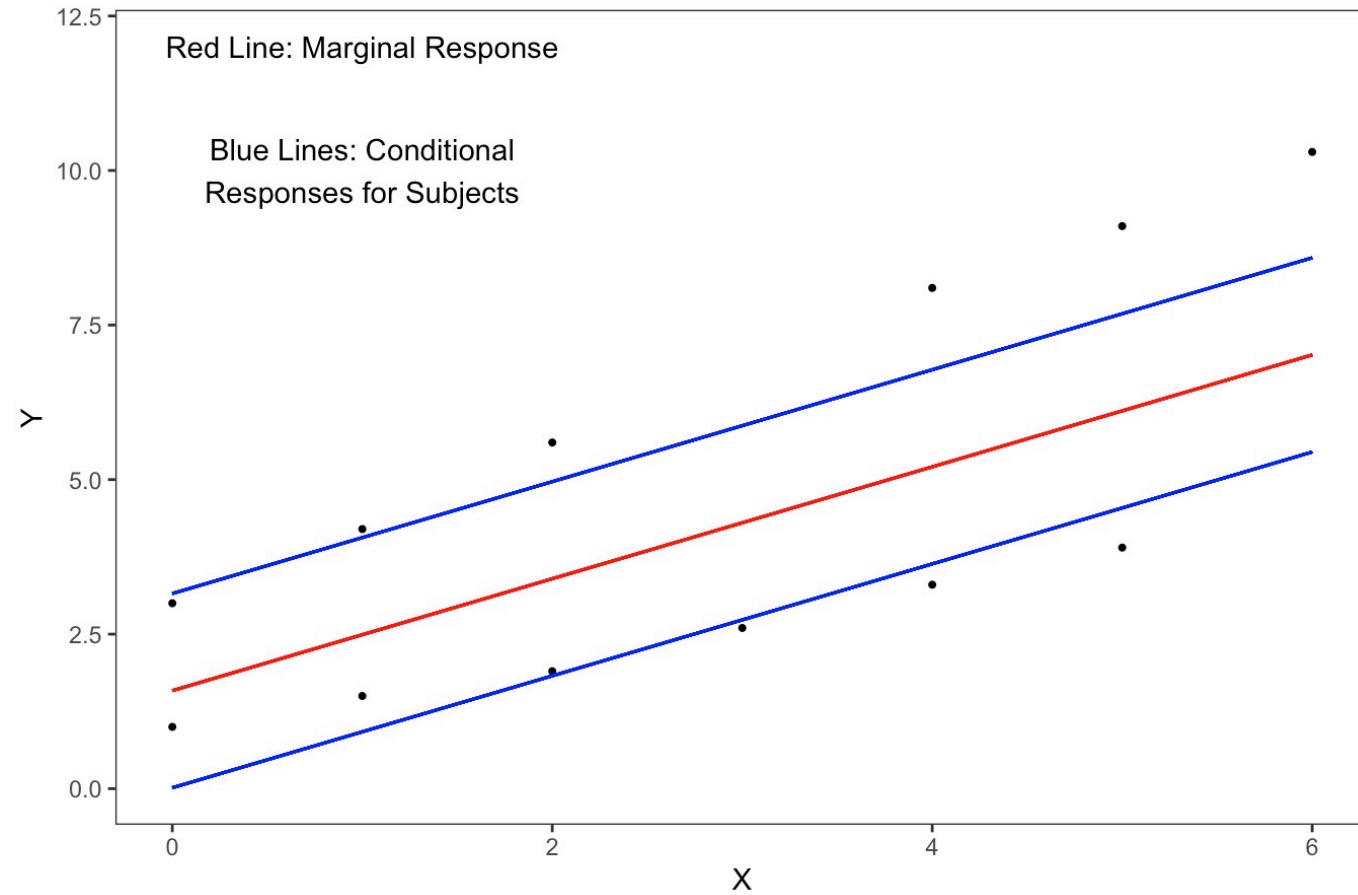
A very simple example involved just a single predictor X :

$$y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{0i} + \varepsilon_{ij} =$$

$$(\beta_0 + b_{0i}) + \beta_1 X_{ij} + \varepsilon_{ij}$$

- This is just like a simple bivariate regression model, except for the random intercept term to account for grouped, potentially correlated observations from the same group or subject

Random intercept model plot



Random intercept model

- Since the β and X terms are assumed fixed, the equation involves only two random terms, b_j and ε_{ij}
- Since b_{0i} and ε_{ij} are assumed to be independent, the variance of each observation is

$$\text{Var}(y_{ij}) = \text{Var}(b_{0i}) + \text{Var}(\varepsilon_{ij}) = \sigma_{b_0}^2 + \sigma_{\varepsilon}^2$$

Random intercept model

- The G matrix is an $n \times n$ diagonal matrix with $\sigma_{b_0}^2$ along the diagonal
- The R matrix is an $N \times N$ diagonal matrix with σ_ε^2 along the diagonal
- The covariance matrix of the y_{ij} is $V = ZGZ' + R$
- V has a “block-diagonal structure” reflecting correlation among observations from the same subject, and independence among observations from different subjects

Random intercept model Z and G matrix example

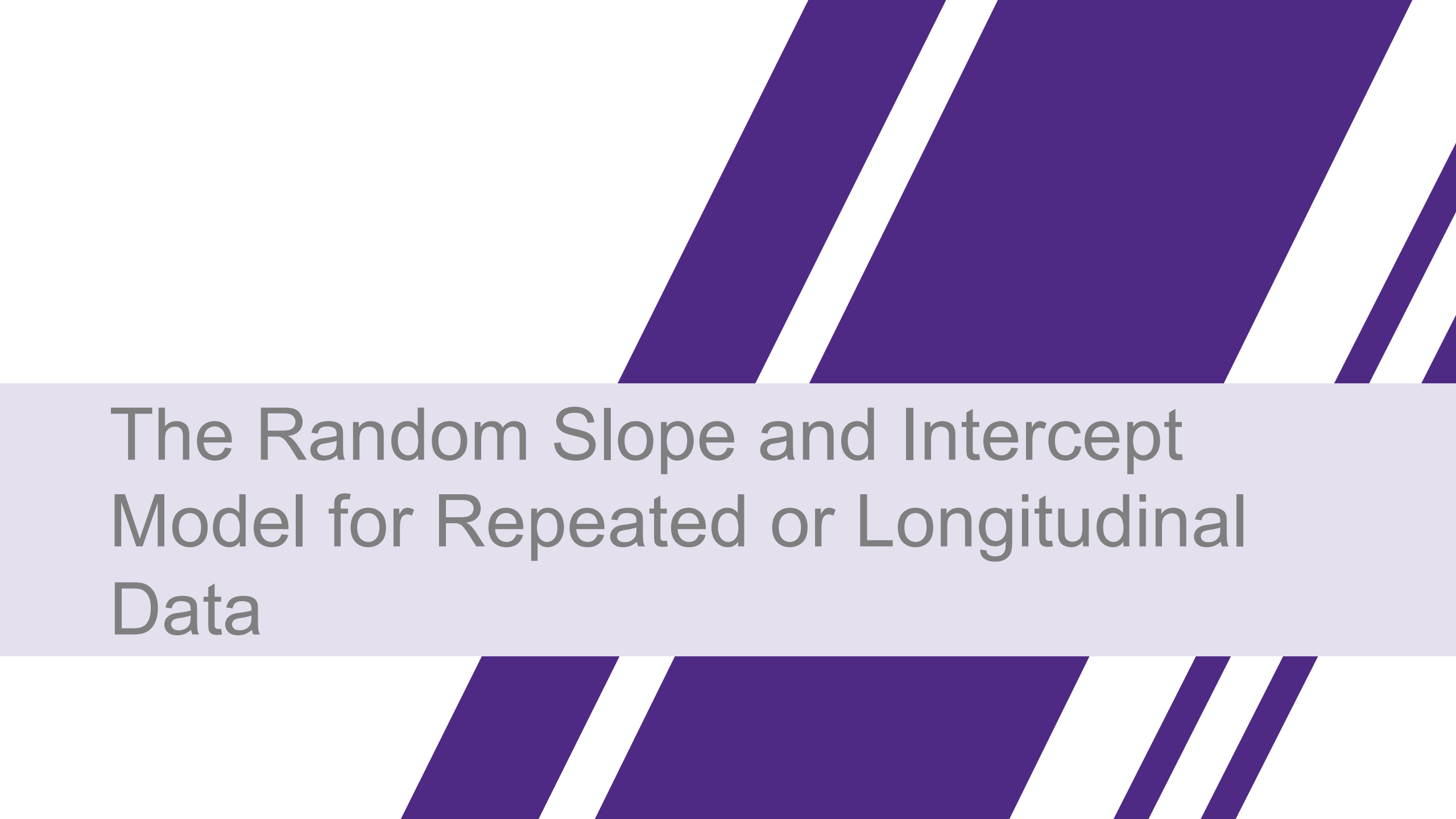
Suppose we had $n = 3$ subjects, with $n_1 = 3$, $n_2 = 2$, and $n_3 = 2$, so $N = 7$. Then

$$Z = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$G = \begin{bmatrix} \sigma_{b_0}^2 & 0 & 0 \\ 0 & \sigma_{b_0}^2 & 0 \\ 0 & 0 & \sigma_{b_0}^2 \end{bmatrix}$$

Random intercept model V matrix example

$$\begin{bmatrix} \sigma_{b_0}^2 + \sigma_\varepsilon^2 & \sigma_{b_0}^2 & \sigma_{b_0}^2 & 0 & 0 & 0 & 0 \\ \sigma_{b_0}^2 & \sigma_{b_0}^2 + \sigma_\varepsilon^2 & \sigma_{b_0}^2 & 0 & 0 & 0 & 0 \\ \sigma_{b_0}^2 & \sigma_{b_0}^2 & \sigma_{b_0}^2 + \sigma_\varepsilon^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{b_0}^2 + \sigma_\varepsilon^2 & \sigma_{b_0}^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{b_0}^2 & \sigma_{b_0}^2 + \sigma_\varepsilon^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{b_0}^2 + \sigma_\varepsilon^2 & \sigma_{b_0}^2 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{b_0}^2 & \sigma_{b_0}^2 + \sigma_\varepsilon^2 \end{bmatrix}$$



The Random Slope and Intercept Model for Repeated or Longitudinal Data

Random slope and intercept model

In this specific model one of the fixed predictors in the X matrix is t_{ij} , reflecting the time of the j^{th} observation within the i^{th} subject, and we also add t_{ij} to the Z matrix to model a random slope b_{1i} for time for each subject, in addition to the random intercept b_{0i} :

$$y_{ij} = \beta_0 + \beta_1 X_{ij1} + \dots + \beta_k X_{ijk} + \beta_t t_{ij} + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij}$$

- Any of the other X matrix terms could also be represented in the Z matrix to add an additional random slope, but for repeated or longitudinal data, often only one random slope term is involved.

Random slope and intercept model

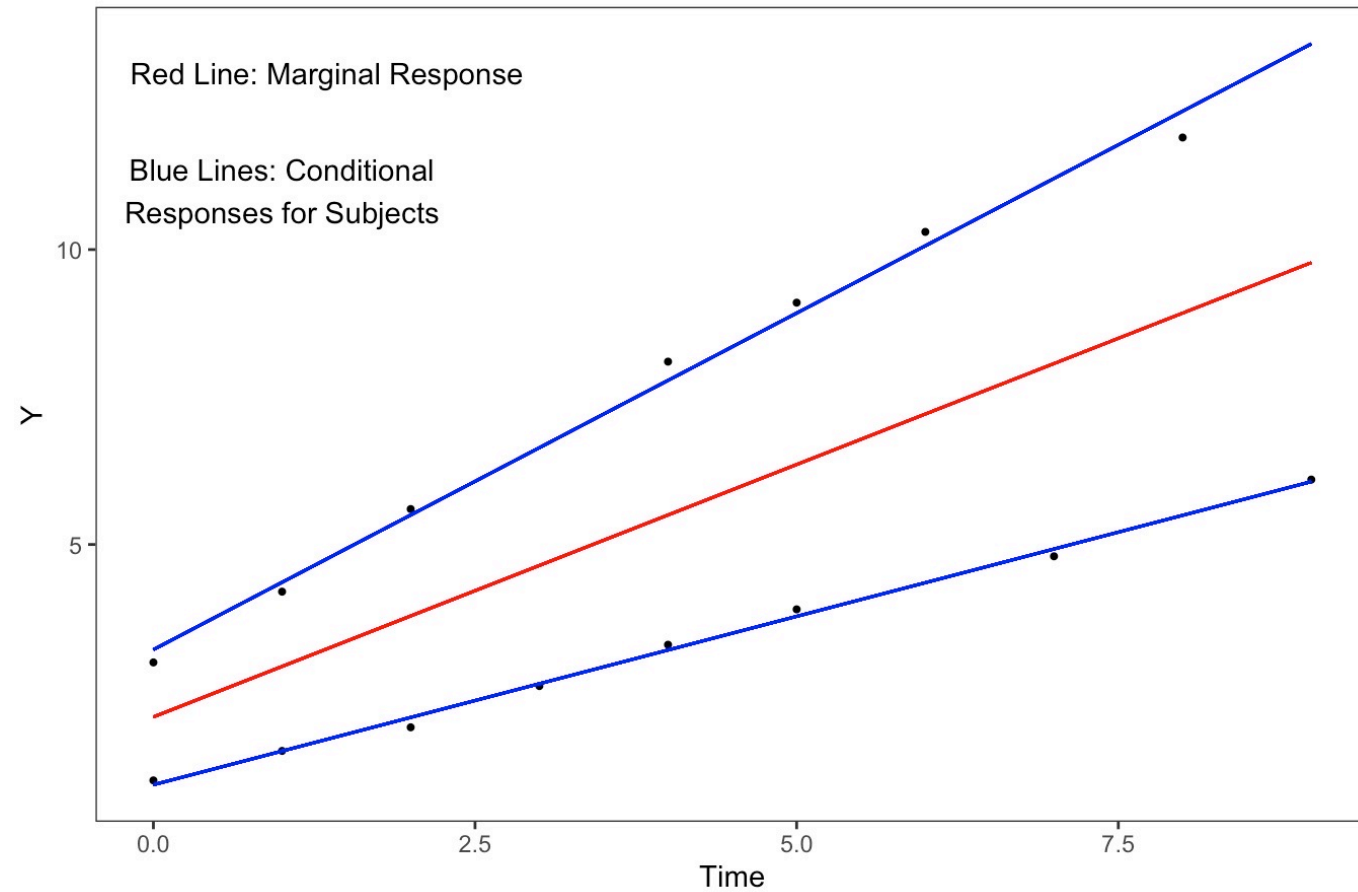
- Note that time here is treated as a continuous variable, not a categorical factor, as is often the case with planned repeated measures at fixed time points.
- Values of time need not be the same for different subjects, and the number of observations per subject need not be the same.
- This allows a great deal of flexibility in modeling observational data and also allows us to handle repeated measures data with missing time points for some subjects without simply dropping entire subjects, as we would have to do with the multivariate general linear model approach to repeated measures.

Random slope and intercept model example

The simplest random slope and intercept model for repeated measures or longitudinal data has only time as a fixed predictor and as a random effect specific to each subject:

$$y_{ij} = \beta_0 + \beta_t t_{ij} + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij} =$$
$$(\beta_0 + b_{0i}) + (\beta_t t_{ij} + b_{1i} t_{ij}) + \varepsilon_{ij}$$

Random slope and intercept model plot



Random slope and intercept model

- The model involves three random terms: b_{0i} , b_{1i} , and ε_{ij}
- b_{0i} , b_{1i} are again assumed independent of ε_{ij}
- It's possible to fit the model assuming independence between b_{0i} and b_{1i} , but typically we allow these to be correlated, resulting in estimation of three variance components in addition to the residual variance σ_ε^2 :
- $\sigma_{b_0}^2$ is again the variance of the random intercept term
- $\sigma_{b_1}^2$ is the variance of the random slope term
- $\sigma_{b_0b_1}$ is the covariance of the random slope and intercept terms

Random slope and intercept model

- Recall that in the random intercept model the variance of each y_{ij} is:

$$\text{Var}(y_{ij}) = \text{Var}(b_{0i}) + \text{Var}(\varepsilon_{ij}) = \sigma_{b_0}^2 + \sigma_{\varepsilon}^2$$

- In the random slope and intercept model, it's much more complicated:

$$\begin{aligned}\text{Var}(y_{ij}) = \text{Var}(b_{0i}) + 2t_{ij}\text{Cov}(b_{0i}, b_{1i}) + t_{ij}^2\text{Var}(b_{1i}) + \text{Var}(\varepsilon_{ij}) = \\ \sigma_{b_0}^2 + 2t_{ij}\sigma_{b_0b_1} + t_{ij}^2\sigma_{b_1}^2 + \sigma_{\varepsilon}^2\end{aligned}$$

Random slope and intercept model

- The G matrix is a $2n \times 2n$ block diagonal matrix with n 2×2 variance-covariance matrices for the random effects along the diagonal:

$$G_i = \begin{bmatrix} \sigma_{b_0}^2 & \sigma_{b_0 b_1} \\ \sigma_{b_0 b_1} & \sigma_{b_1}^2 \end{bmatrix}$$

- The R matrix remains an $N \times N$ diagonal matrix with σ_ε^2 along the diagonal
- The covariance matrix of the y_{ij} remains $V = ZGZ' + R$

Random slope and intercept model Z and G matrix example

Suppose we had $n = 2$ subjects, with $n_1 = 3$ and $n_2 = 2$, so $N = 5$.
Then

$$Z = \begin{bmatrix} 1 & t_{11} & 0 & 0 \\ 1 & t_{12} & 0 & 0 \\ 1 & t_{13} & 0 & 0 \\ 0 & 0 & 1 & t_{21} \\ 0 & 0 & 1 & t_{22} \end{bmatrix}$$

$$G = \begin{bmatrix} \sigma_{b_0}^2 & \sigma_{b_0 b_1} & 0 & 0 \\ \sigma_{b_0 b_1} & \sigma_{b_1}^2 & 0 & 0 \\ 0 & 0 & \sigma_{b_0}^2 & \sigma_{b_0 b_1} \\ 0 & 0 & \sigma_{b_0 b_1} & \sigma_{b_1}^2 \end{bmatrix}$$

Random slope and intercept model V matrix structure example

The structure of V remains block diagonal, with observations from the same subject correlated, but observations from different subjects independent:

$$V_{structure} = \begin{bmatrix} V_{11} & C_{11,12} & C_{11,13} & 0 & 0 \\ C_{11,12} & V_{12} & C_{12,13} & 0 & 0 \\ C_{11,13} & C_{12,13} & V_{13} & 0 & 0 \\ 0 & 0 & 0 & V_{21} & C_{21,22} \\ 0 & 0 & 0 & C_{21,22} & V_{22} \end{bmatrix}$$

Random slope and intercept model V matrix example

Due to space issues, here I show only the 3×3 submatrix associated with the first subject, V_1 :

$$\begin{bmatrix} \sigma_{b_0}^2 + 2t_{11}\sigma_{b_0b_1} + t_{11}^2\sigma_{b_1}^2 + \sigma_\varepsilon^2 & \sigma_{b_0}^2 + (t_{11} + t_{12})\sigma_{b_0b_1} + t_{11}t_{12}\sigma_{b_1}^2 + \sigma_\varepsilon^2 & \sigma_{b_0}^2 + (t_{11} + t_{13})\sigma_{b_0b_1} + t_{11}t_{13}\sigma_{b_1}^2 + \sigma_\varepsilon^2 \\ \sigma_{b_0}^2 + (t_{11} + t_{12})\sigma_{b_0b_1} + t_{11}t_{12}\sigma_{b_1}^2 + \sigma_\varepsilon^2 & \sigma_{b_0}^2 + 2t_{12}\sigma_{b_0b_1} + t_{12}^2\sigma_{b_1}^2 + \sigma_\varepsilon^2 & \sigma_{b_0}^2 + (t_{12} + t_{13})\sigma_{b_0b_1} + t_{12}t_{13}\sigma_{b_1}^2 + \sigma_\varepsilon^2 \\ \sigma_{b_0}^2 + (t_{11} + t_{13})\sigma_{b_0b_1} + t_{11}t_{13}\sigma_{b_1}^2 + \sigma_\varepsilon^2 & \sigma_{b_0}^2 + (t_{12} + t_{13})\sigma_{b_0b_1} + t_{12}t_{13}\sigma_{b_1}^2 + \sigma_\varepsilon^2 & \sigma_{b_0}^2 + 2t_{13}\sigma_{b_0b_1} + t_{13}^2\sigma_{b_1}^2 + \sigma_\varepsilon^2 \end{bmatrix}$$

- Note that all variances and covariances are functions of the specific time points for the observations, in addition to the variance components for the random effects.



Some Complexities of Mixed Models

Mixed model complexities

The power and flexibility of mixed models are accompanied by some important and substantial complexities

- Unlike the general linear model, estimation of mixed model equations requires iterative techniques based on likelihood methods: maximum likelihood (ML) or restricted/residual maximum likelihood (REML)
- Such models may take much longer to estimate than those handled via least squares methods
- Convergence to stable and globally correct values of parameter estimates can sometimes be difficult or impossible to achieve in cases where the model is not consistent with the data or the data are sparse

Mixed model complexities

- Sample sizes generally need to be substantially larger in mixed models in order to obtain good results
- Model selection is more complicated than with only fixed effects
- Denominator degrees of freedom associated with F statistics in unbalanced mixed models are not simple values and have to be estimated
- Some controversy exists about the validity of the assumed null distributions for these statistics



Software for Linear Mixed Models

Primary software options for linear mixed models

- All of the major general purpose statistical software packages available at Northwestern offer mixed models capabilities
- The most popular package in R for such methods is lme4 and its lmer function (typically supplemented with lmerTest)
- SAS has a MIXED procedure that offers linear mixed models in a very general framework with a huge array of options
- SPSS has a MIXED procedure in a similar general framework with somewhat fewer options
- Stata offers a mixed procedure for mixed/hierarchical/multilevel linear models



A Demo with R



Questions?