Introduction to Linear Mixed Models: Modeling continuous longitudinal outcomes

Dr Cameron Hurst
cphurst@gmail.com

CEU, ACRO and DAMASAC, Khon Kaen University

4th Febuary, 2557

Some motivational datasets

Before we start, I would like to introduce some datasets that will set the scene for this session(correlated continuous outcomes) and the next session (correlated categorical outcomes). All three datasets come with libraries used for longitudinal data modeling:

1. Sleep study (`lme4`)
2. Autism data (`WWGbook`)
3. Respiratory data (`geepack`)

The first two will be used to demonstrate linear mixed models(LMMs), and the third dataset (which has a binary outcome) will be used in the next session to consider Generalized Linear Mixed Models(GLMMs) and Generalized Esitmating Equations(GEEs).

Sleep data

This study, conducted by Belensky(2003), investigates the effect of sleep deprivation on reaction times. On day 0 subjects had normal amounts of sleep, and starting that night, had 3 hours sleep per night. Average reaction time was measured across a series of tests administered each day, to each subject. Variables are:

- **Reaction:** Average reaction time (**continuous outcome**)
- **Day:** Number of days of sleep deprivation (day: 0,1,2,...,9)
- **Subject:** Subject ID

Questions for later:

1. Which **factors** is **FIXED** and which is **RANDOM**?
2. Is the single fixed effect a **WITHIN**-subject, or **BETWEEN**-subject effect?

Autism data

This study (Oti et al, 2006) investigates the effect of the level of communication development (as classified at age 2) on social development in Autistic children. Cohort participants are initially measured at 2 years old and then followed up until age of 13:

- **VSAE:** parent-reported Vineland Socialization Age Equivalent
- **Age:** Age in years (2, 3, 5, 9, 13)
- **Sicdegp:** Expressive language score at 2yo:Low, Med, High
- **Childid:** Unique child identifier

Now:

1. Of the three factors, which are Fixed and which is Random?
2. For each fixed factor: **Within-** or **Between-** subject?

Respiratory data(for next session)

A multi-centre, placebo-controlled RCT to investigate the efficacy of a 'drug' on respiratory illness. A group of 111 patients (from two centres) were randomized to either the placebo or treatment arm. Respiratory illness (y/n) was observed at baseline, and then again on three subsequent visits. Variables:

- **Respiratory illness** present (y/n);
- **Visit:** 1 (baseline) and three follow-up visits (2,3 and 4);
- **Treat:** P=Placebo or A=Active
- **Patient.id:** Unique patient identifier
- **Centre.id:** Centre ID (1 or 2)

Possible confounding effects: The factor Gender and covariate Age at baseline

Respiratory data(for next session)

Now for respiratory data:

1. How does the outcome differ (compared to Sleep and Autism studies)?
2. Of the many **EFFECTS**, which are **FIXED** and which are **RANDOM**?
3. For each **FIXED** effect, which is **WITHIN-** and which is **BETWEEN-**subject

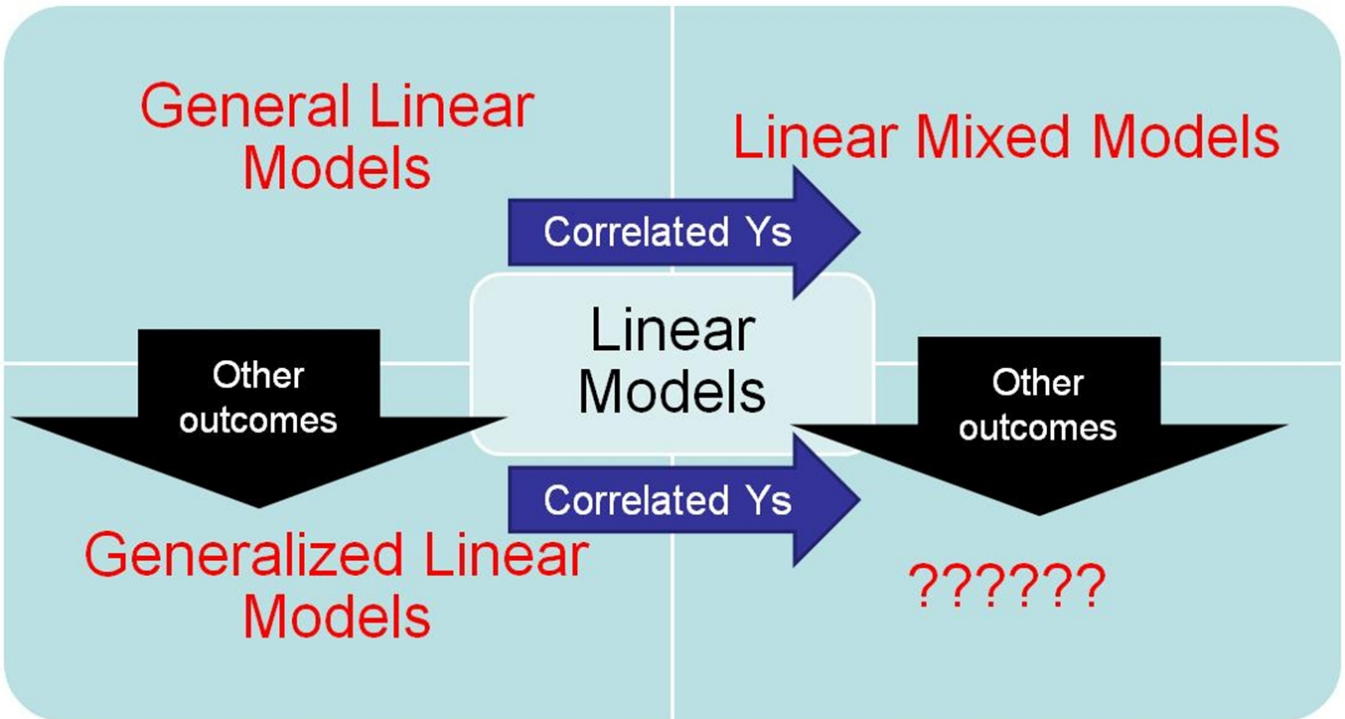
BONUS POINTS: What is the name of the rather simplistic 'classical' bivariate tests that has been traditional used for binary outcomes for pre-post studies (similar to this)? *Hint: Only works for 2 x 2 data*

What we will cover....

1. Introduction
   - Longitudinal data
   - Fixed and Random effects
2. Methods for the analysis of (continuous) longitudinal data
   - 'Classical' approaches
   - Linear Mixed Models
   - Worked example of a Linear Mixed Model in R
3. Sample size calculations for LMMs

Overview of linear models

How do linear models relate??



Repeated measures data

- Most linear models we use in biostatistics are for the analysis of cross-sectional (uncorrelated) data
- What about if our sampling units are measured repeatedly over time?
- Data sets containing such repeated measurements are called **repeated measures**, **time-course**, or **longitudinal** datasets
- Studies involving such data are often called **longitudinal** studies, and represent a TYPE of cohort study

Warning: Terminology

Take care of the word "**cohort**". By definition it just means exposure preceeds outcome, and such studies MAY, or MAY NOT involve repeated measurements

Repeated measures data

What difference does it make?

- What's the harm if we just use our standard cross-sectional regression methods (general linear models, generalized linear models etc.)?
- What assumption do all these 'cross-sectional' methods share that would be violated?
- **ANS: Independence:**
  - **Each (X, Y) observation is independent of every other (X, Y) observation**
  - **That is, (X_i, Y_i) is independent of (X_j, Y_j); for all i, j where i $\neq$ j**

Repeated measures data

Still....So what??

- What's the problem with 'correlated' observations?
- We can assume measurements taken for a single subject (i.e. WITHIN-subject) are likely to be more similar then observations among subjects (BETWEEN-subjects)
- Lower variation for within-subject observations should be considered the advantage it is (if modelled properly)
- BUT problem with cross-sectional methods is that they assume all variation is 'between-subject' (so can underestimate variation in the data)
- SO, if we use a standard cross sectional approach, variance estimates will be an under estimate $\Rightarrow$ type I errors

Think about this. WHY?

Advantages of repeated measures data

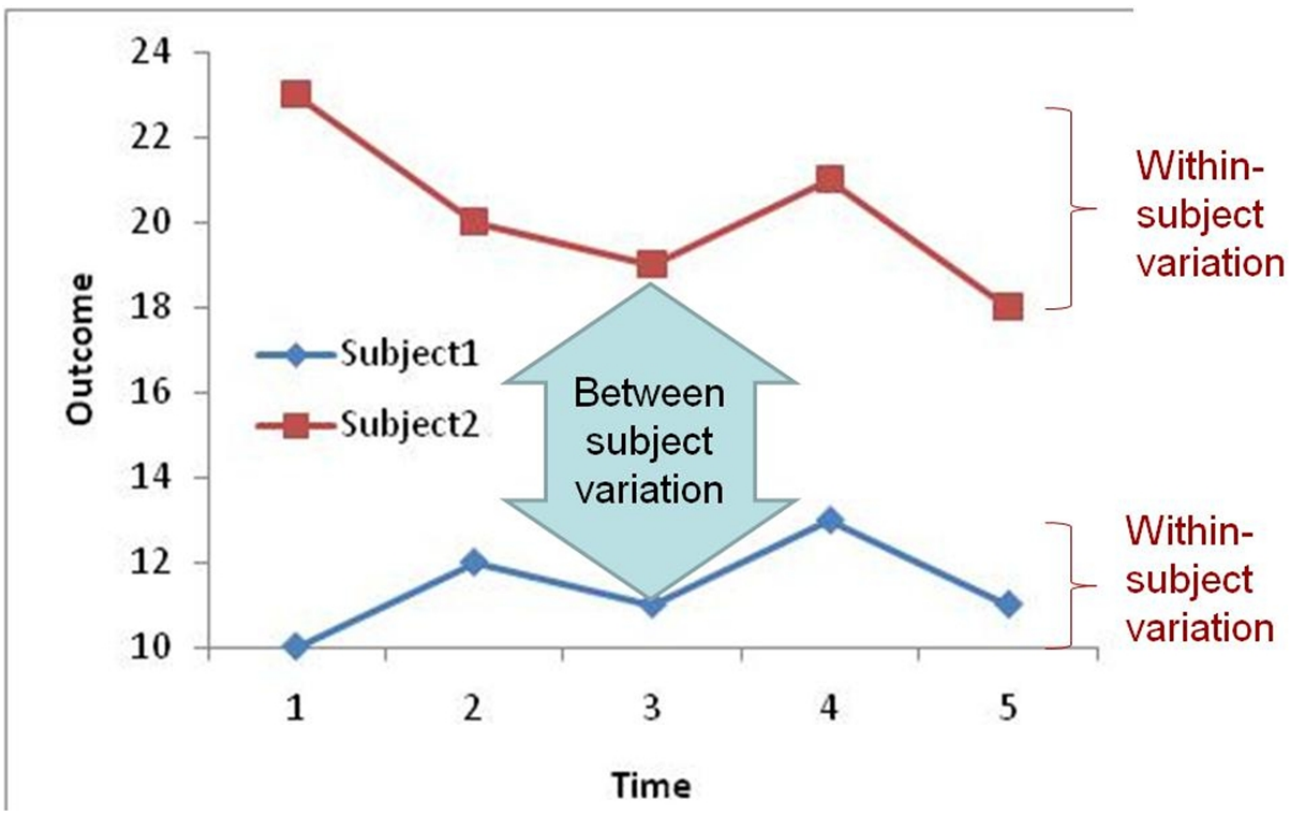
- Remember that we are trying to reduce error and minimize bias (E.g. adjusting for other covariates)
- In repeated measures, individuals are considered as 'blocks' of observations
- Assume within-subject variation lower than among subjects
- Advantage - can conduct complex designs with fewer observational (or experimental) units

Between vs Within subject variation

- Between-subject variation is how we would expect differences between subjects (or groups thereof) to manifest itself (this is the type of variation we are used to) - e.g. A standard ANOVA (General Linear Model)
- Within-subject variation is how a **given** individual varies over time

Between vs Within subject variation

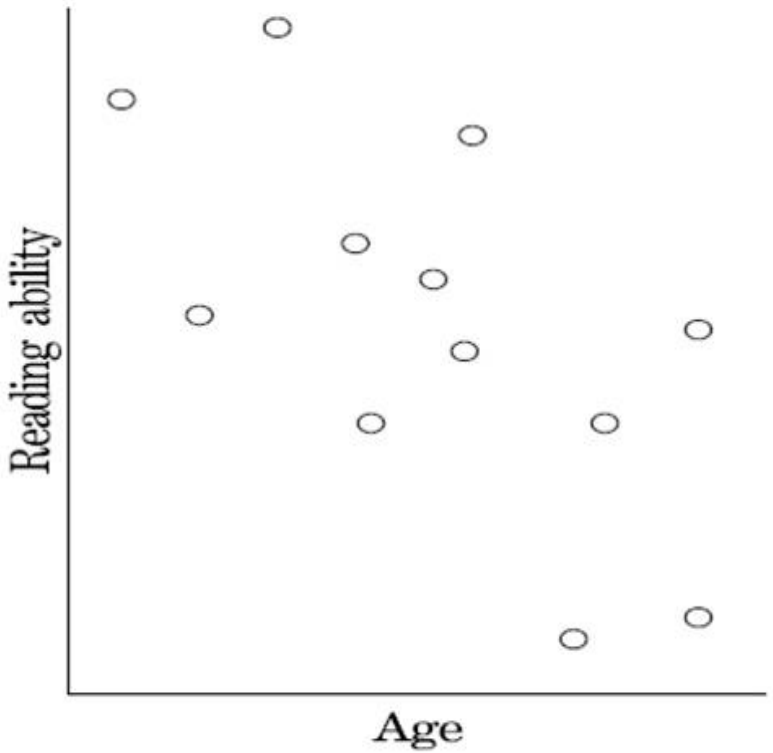
Figure : Within Vs Between variation for two individuals



Repeated measures data

Not accounting for repeated measurement

- Cross-sectional data (observations are independent)
- Greater age ⇒ Poorer ability

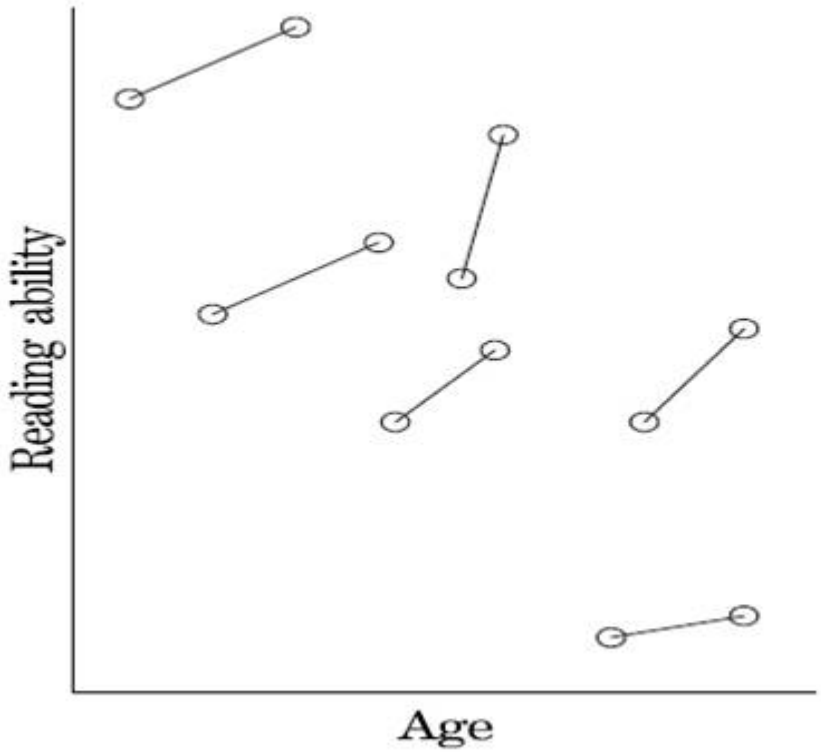


Source: Diggle et al., Analysis of Longitudinal Data

Repeated measures data

Accounting for repeated measurement

- Longitudinal data (observations are linked)
- Greater age ⇒ Better ability



Source: Diggle et al., Analysis of Longitudinal Data

Other types of correlated data

Clustered data

- We can think of longitudinal data as repeated measures over time.
- Clustered data is repeated measures over space
  - E.g. If we consider a multi-centre study involving a number of clinics we would expect people attending same clinics to be more similar (relative to those attending other clinics).
  - We need to account for this in modelling
- Good news is that methods used for clustered data are identical to those used for longitudinal data
- Clustered data turns up in health studies all the time. E.G.
  - Ophthalmology: Right and left eyes of the same patient
  - Multi-center clinical trials

Within-subject correlation

- Between variable correlation considers the association between two variables (e.g. Exercise and depression levels) - this is the correlation we are used to thinking about
- Within-subject correlation considers the correlations between observations considered at two times (for each subject)
  - For example, White blood cell count before and after a treatment
  - Those with higher before-treatment WBC would be expected to have higher WBC after treatment (relative to other subjects)
- As you would expect, within-subject correlation is generally positive

Random effects

- Random effects are those where the particular 'groups' (levels) are drawn randomly from a population of groups.
- For example, if have 10 patients drawn from a population on which we take five (repeated) measurements.
- It doesn't make sense to test:
  $H_0 : \mu_{subject1} = \mu_{subject2} = \cdots = \mu_{subject10}$
- If we rejected this hypothesis, we can't make any meaningful inference about the differences between the specific patients (to the population)
- If we were going to tests a hypothesis about this random effect (Patient), which we rarely do, we would test:
  $H_0 : \sigma^2_{patients} = 0$ (patients in popn. don't vary)

Your turn: do we understand differences between fixed and random effects?

Fixed effects are the ones we know (and love), where we would expect a differences between two groups to be **FIXED** as we move from the sample to the population

- Which (below) are **Fixed-** and which **Random-**effects??
- **AND** for fixed effects (only), which are **within-subject** and which **between-subject**?
  1. New treatment vs Placebo control
  2. Patient (A, B, C, D, E, F or G)
  3. Visit (Baseline, 3 month, 6 month and 1 year)
  4. Brand of toothpaste (A, B or C) $\rightarrow$ Tricky
  5. Day of year (any possible day of the year)
  6. Type of Cancer

What about datasets at beginning of lecture??

Fixed and random effects

- Generally only hypotheses about fixed effects interest us.
- We often just want to account for (deal with) the random effect in the model (testing a hypothesis concerning this variation doesn't tell us anything useful)
- Problem with random effects is we often need a lot of model parameters (i.e. 40 subjects $\geq$ 40 parameters)
- Now we understand the purpose of fixed and random effects, we can start to consider the different methods (models) for the analysis of repeated measure data

Methods for longitudinal continuous outcomes

We will discuss four methods for the analysis of continuous longitudinal outcomes:

1. Repeated measures ANOVA (RM-ANOVA)
2. Repeated measures multivariate ANOVA (RM-MANOVA)
3. Linear mixed models (LMM)
4. Linear marginal models (Next session)

but we will only cover the last two of these in any detail

Classical approaches

- The first two of these techniques have fallen out of favour recently (as they make restrictive, and often, unrealistic assumptions about the data, or they are not practical)
- But I will cover them briefly
- Why bother? They illustrate useful insights into repeated measures data and its modelling
- Also their models can be represented (very closely) in a Linear marginal model anyway (so we can directly compare and assess their adequacy) $\rightarrow$ More next session

Assumptions of RM-ANOVA

Beyond those assumptions normally associated with an ANOVA, the RM-ANOVA makes more:
Compound symmetry

1. Constant variance $\rightarrow \sigma^2(Y_{ij}) = \sigma_e^2 + \sigma_b^2$
2. Constant covariance $\rightarrow Cov(Y_{ij}, Y_{ik})$

The second of these states that regardless of how far apart two repeated measures are, they will be just as correlated
e.g Corr(Day1, Day2) = Corr(Day1, Day20)

Compound symmetry is a type of **residual covariance structure**
(We cover covariance structures in MUCH more detail when we get to Marginal Models).

Why not RM-ANOVA?

- The compound symmetry (covariance structure) assumption associated with the Repeated measures ANOVA renders it too unrealistic in many health-based longitudinal studies.
- Besides, many stats packages (esp. SPSS) tend to make the use of RM-ANOVAs quite painful
- We will only implement the RM-ANOVA by using the much more flexible Linear Marginal Model (next session)
  - i.e. RM-ANOVA $\rightarrow$ Linear Marginal Model with Compound symmetry residual covariance structure

Repeated measure MANOVA

- RM-MANOVA (Repeated Measures **Multivariate** Analysis of variance) represents an alternative approach to the analysis of repeated measures data
- More computationally intensive than RM-ANOVA, but it relaxes some of the (often unrealistic) assumptions of the RM-ANOVA in regards to the residual covariance structure.
- In the standard MANOVA approach, no particular structure is assumed in the residual covariance structure (i.e. The covariance structure is called **unstructured**)
- Again, we can implement the RM-MANOVA using the Linear Marginal Model
  - i.e. RM-MANOVA $\approx$ Linear Marginal Model with Unstructured residual covariance structure

Why we won't use classical approaches

- Generally, RM-ANOVAs and RM-MANOVAs presents major restrictions that make their use impractical.
- RM-ANOVAs assumptions about a compound symmetric covariance structure are too simplistic (i.e. often invalid) in most health studies.
- RM-MANOVA needs a large sample sizes in order to be useful, also has a major problem with missing values
- Both methods have major problems with missing values or unbalanced data:
  - They can't deal with different numbers of repeated measures for different subjects
  - If even one value is missing for a subject, that subject has to be excluded (list-wise deletion)

Linear Mixed Models for longitudinal data

This leaves the last two methods we will talk about:

1. Linear **MIXED** models mix (consider) both **fixed** and **random** effects (Hence the name 'Mixed' models)
2. Linear **MARGINAL** models represent an alternative approach accounting for correlation of observations through the use of **Residual covariance structures**

Keeping it simple

When we cover the comparitively complex conditional and marginal models, remember one thing: **These models are a way of dealing with correlated data so we can consider it the same way as we consider independant data:** Within-subject effects can be treated as (standard) between-subject effects

Linear Mixed Model: A general representation

The Linear Mixed (Effects) model can be represented:

$$Y_{ti} = \beta_0 + \beta_1 X_{ti}^{(1)} + \beta_2 X_{ti}^{(2)} + \cdots + \beta_{k-1} X_{ti}^{(k-1)} + u_{0i} + u_{1i} Z_{ti}^{(1)} + \cdots + u_{qi} Z_{ti}^{(q)} + \epsilon_{ti} \quad (1)$$

The value of t ($t = 1, 2, \cdots, n_i$) indexes the longitudinal observations for each subject i ($i = 1, 2, \cdots, m$) over time. Using matrix notation:

$$Y_i = X_i \beta + Z_i u_i + \epsilon_i$$

The first part of the model, containing $X_i \beta$, is the **fixed effect(s)** component of the model (a standard general linear model), whereas the $Z_i u_i$, represents the **random effect(s)**.

LMMs: Defining the terms

In the model:

$$Y_i = X_i\beta + Z_i u_i + \epsilon_i$$

Y_i is a vector of continuous responses for the i^{th} subject
X_i is an $n_i \times k$ design matrix for the corresponding fixed effect covariate values for each of the n_i observations collected on the i^{th} subject
β is a vector of regression coefficients for the fixed effects, X
Z_i is the $n_i \times q$ design matrix that represents the known values of the q covariates, $Z^{(1)}, \cdots, Z^{(q)}$, for the i^{th} subject
u_i is a vector of q random effects (more later), with $u_i \sim N(0, D)$
and ϵ_i is the residual, with $\epsilon_i \sim N(0, R)$
Next we will define the matrices, D an R

The matrix D

- Elements along the main diagonal of the D matrix represent the variances of each random effect in u_i, and the off-diagonal elements represent the covariances between two corresponding random effects
- We will discuss elements of the D matrix as we discuss different LMM models
- In particular, this matrix is used to define the random intercept and random coefficient models

The matrix R

- The R matrix, called the **residual (or error) covariance structure** is a key matrix in Marginal models
- First, unlike standard linear models, it allows errors (and therefore observations) to be correlated to each other
- Second,there are several ways of specifying R to make our model better fit the nature of our data
- Later (when we discuss marginal models) we will consider 4 ways of specifying R:
  1. Independent residual covariance structure
  2. Compound symmetry residual covariance structure
  3. Autoregressive(1) residual covariance structure
  4. Unstructured residual covariance structure

Key differences between a General (normal) linear Model and the LMM

- One way of understanding the LMM better is to contrast it against the General Linear Model
- We can view the General Linear Model as an example of a LMM with:

$Z = 0$ i.e. No random effects; and

$R = \sigma^2 I$ i.e. Independent observations

Keeping it simple: Linear models, Mixed Models and Marginal models

A standard general linear model:

$$y_{ij} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \epsilon_{ij}$$

To deal with the correlated data, the linear mixed model and linear marginal models...

$$y_{ij} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \text{Put something here}$$

The 'something' is quite different between the **mixed** model approach (which is a 'subject-specific' or **conditional** model), and the **marginal** model approach (a **population-averaged** approach)

Keeping it simple: Mixed vs Marginal models

For **mixed** (aka conditional, or subject-specific) models, the correlated data is dealt with EXPLICITLY in the model (i.e. terms are added to deal with random effects)

$$y_{ij} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \text{random effects} + \epsilon_{ij}$$

IN CONTRAST, for **marginal** (aka population averaged) models, the correlated nature of the data is dealt with in the residual term ϵ

$$y_{ij} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0, R)$ where R is the residual covariance structure (and FULLY reflects correlations among observation).

Our first LMM: Random intercept model

- Let's start with a quite simple LMM, the **Random intercept model**
- This model allows each subject(after accounting for the fixed effects) to vary at baseline
- But we should note that the effect of time, is assumed to be the same for all individuals (all individuals share a common slope over time)

Random intercept model specification

Taking the (general) linear mixed model specified above:

$$Y_i = X_i\beta + Z_i u_i + \epsilon_i$$

In the random intercept model, the Z matrix (as there are only intercepts) becomes a column of ones, so the model simplifies down to:

$$y_{ij} = X_{ij}\beta + u_i + \epsilon_{ij}$$

Now defining the terms.....

Our first LMM: Random intercept model

The Random intercept model:

$$y_{ij} = X_{ij}\beta + u_i + \epsilon_{ij}$$

Where
y_{ij} is the the j^{th} repeated measure on the i^{th} subject
X_{ij} is the corresponding vector of fixed effect covariates/factors
β is the corresponding fixed-effect regression parameters (up to this point, a standard general linear model)
u_i is the (random) effect due to subject; The change (from average) intercept is subject i
ϵ_{ij} is the residual (error not explained by covariates); with
$\epsilon_{ij} \sim N(0, \sigma_e^2)$
$u_i \sim N(0, \sigma_u^2)$ is the variation among subjects not accounted for by covariates in X

Simplifying the random intercept model

If we do a little algebraic rearrangement (as we did with general linear models a few sessions back):

$$y_{ij} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + u_i + \epsilon_{ij}$$

Can be rearranged to:

$$y_{ij} = y_{ij} = (\beta_0 + u_i) + \beta_1 X_1 + \beta_2 X_2 + \cdots + \epsilon_{ij}$$

So we can see u_i is the **difference from the average intercept, β_0, due to patient i**

Sleep data

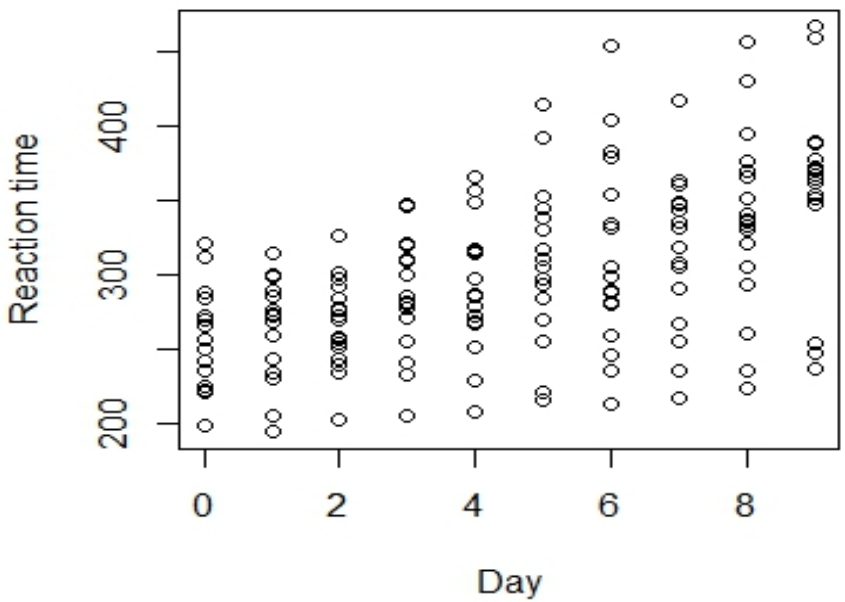
Recall Belensky(2003) sleep study. We want to investigates the effect of sleep deprivation on reaction times. On day 0 subjects had the normal amount of sleep, and starting that night, had 3 hours sleep per night. Average reaction time was recorded each day, for each subject. Variables are:

- **Reaction:** Average reaction time (**continuous outcome**)
- **Day:** Number of days of sleep deprivation (day: 0,1,2,...,9)
- **Subject:** Subject ID

We will start by running a **Random Intercept** Linear Mixed Model

Step 1: Eyeball the data

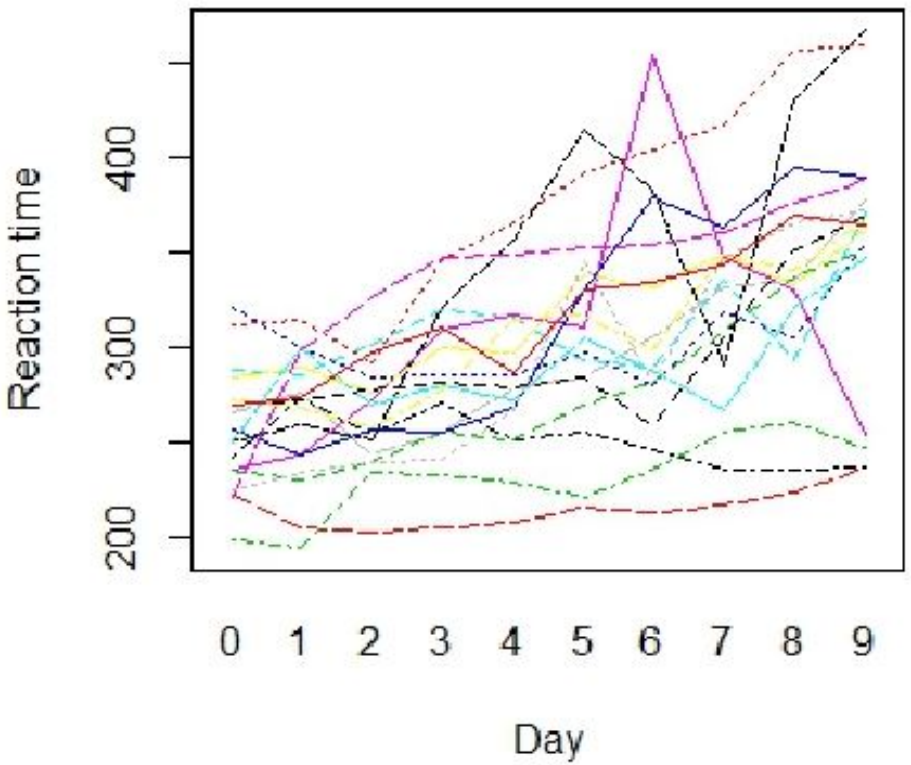
Let's start with a standard scatter plot (where there is no linking of within-subject observations)



Helpful?? Not really.

Step 1: Eyeball the data

Now let's link the within-subject observations



Much better. Note for later. Do you think the subjects have the same slope (i.e. the same response to sleep deprivation)?

Step 2: Specify the model

Now let's state the model in terms of our problem:

$$Reaction_{ij} = \beta_0 + \beta_{Day} Day + Subject_i + \epsilon_{ij}$$

Rearranging:

$$Reaction_{ij} = (\beta_0 + Subject_i) + \beta_{Day} Day + \epsilon_{ij}$$

$Reaction_{ij}$ is the reaction time of the *ith* subject on the *jth* day
β_0 is the overall (average) y-intercept
$Subject_i$ is the CHANGE from β_0 due to being subject i
β_{Day} is the slope due to day (expected change in reaction time due to an extra day of sleep deprivation)
ϵ_{ij} is the random (unexplained) error associated with $Reaction_{ij}$

Step 3: Run the model

- Several libraries in R run LMMs. I prefer `lme4` library
- This doesn't come with the base R (need to download)
- Assume our data already read in and stored in `sleep.df`

R syntax: Random intercept LMM

```
library(lme4)

# Need a null (nothing in it) model first
null.mod<-lmer(Reaction ~ 1 + (1|Subject), data=sleep.df)

# Now fit our model
rint.mod<-lmer(Reaction ~ Days + (1|Subject), data=sleep.df)

# Does model represent an improvement over null
anova(null.mod, rint.mod)

summary(rint.mod)
```

Output 1a) ANOVA

```
> rint.mod<-lmer(Reaction~Days + (1|Subject), data=sleep.df)
> anova(null.mod, rint.mod)
Data: sleep.df
Models:
null.mod: Reaction ~ 1 + (1 | Subject)
rint.mod: Reaction ~ Days + (1 | Subject)
          Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
null.mod   3 1916.6 1926.2 -955.28
rint.mod   4 1802.1 1814.9 -897.05 116.47      1  < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1
```

Output 1b) Coefficients

```
> summary(rint.mod)
Linear mixed model fit by REML
Formula: Reaction ~ Days + (1 | Subject)
   Data: sleep.df
  AIC  BIC logLik deviance REMLdev
 1794 1807 -893.2     1794    1786
Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 1378.18  37.124
 Residual              960.46  30.991
Number of obs: 180, groups: Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept) 251.4051     9.7459   25.80
Days         10.4673     0.8042   13.02
```

Random intercepts model: Interpretation

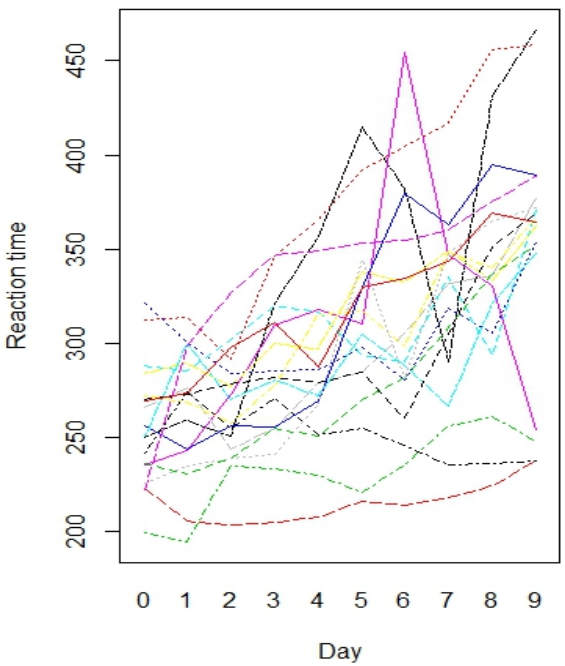
- We can see from output 1a that when we compare the model with the fixed effect *Day* to the null model, there is a significant improvement ($\chi^2_{LR} = 116.47$, $df = 1$, $p < 0.001$)
- When we go to the coefficients table we see $\beta_{Day} = 10.46$, that is, for every extra day of sleep deprivation, we would expect reaction time to increase by 10.46 ms
- You will notice there is no R^2 (this is not a linear regression). Instead we are left with the MUCH LESS friendly **AIC = 1802.1**
- The AIC for the null model (see output 1a) is **AIC = 1910**, that our random intercept model is considerably lower suggest our model is a considerably better fit (remember **lower AIC means better model**)

Random intercepts model: Interpretation

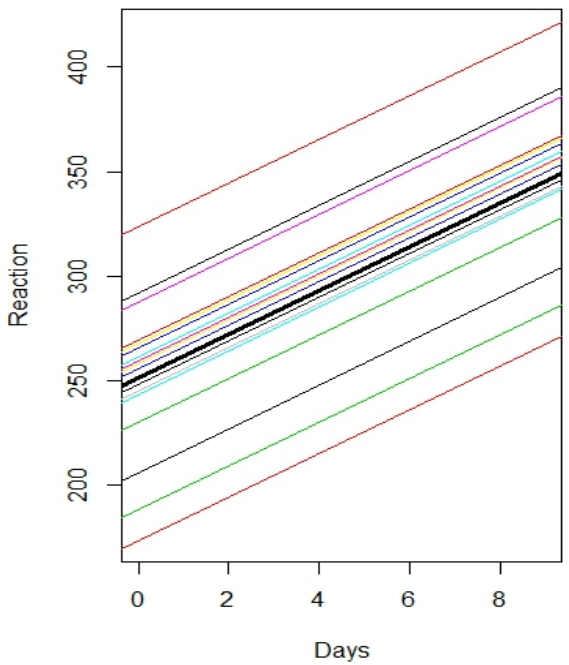
- The average intercept was $\beta_0 = 251.4$ (see Output 1b)
- This implies the AVERAGE reaction time at baseline (after no sleep deprivation) was 251.4 ms
- Our **Random intercept** model allows individuals to vary from this average intercept
- Let's look at how this intercept varied across subjects...

Random intercept model: Adequacy

BUT does this model represent a realistic fit of the data?



Real data



Fit (Random intercepts)

Linear Mixed Model: Random coefficients model

- Perusing the graphs on the previous page suggests that the model doesn't fit the data well.
- It appears the assumption that all subjects have the same slope (the same response to progressive sleep deprivation) is unrealistic
- A better model may be the random coefficients model. This allows subjects to have BOTH their own intercept, and their own slope...

Back to Step 2: Model (Re)specification

The random coefficients model (in our case), will be

$$Reaction_{ij} = \beta_0 + \beta_{Day} Day + Subject_{i0} + Subject_{Days,i} Days + \epsilon_{ij}$$

Rearranging:

$$Reaction_{ij} = (\beta_0 + Subject_{0i}) + (\beta_{Day} + Subject_{Days,i}) Day + \epsilon_{ij}$$

Let's define each of the terms

Back to Step 2: Model (Re)specification

$$Reaction_{ij} = (\beta_0 + Subject_{0i}) + (\beta_{Day} + Subject_{Days,i})Day + \epsilon_{ij}$$

$Reaction_{ij}$ as before
β_0 is the overall (average) y-intercept
$Subject_{0i}$ is the change in y-intercept due to subject i
β_{Day} is overall (average) slope for day
$Subject_{Day,i}$ is change in slope due to subject i

So we can see $Subject_{Day,i}$ as the 'effect modification' of days of sleep deprivation associated with subject i'

Step 2a Model (re)specification: Random coefficients model

R syntax: Random coefficient LMM

```
# Fit random coefficients model
rcoef.mod<-lmer(Reaction ~ Days + (Days|Subject),
data=sleep.df)
# Is it an improvement on the random intercepts model)
anova(rint.mod, rcoef.mod)

summary(rcoef.mod)
```

Output 2a) ANOVA

```
> rcoef.mod<-lmer(Reaction~Days + (Days|Subject), data=sleep.df)
> #Signicant improvment over random intercept?
> anova(ran.int, ran.coef)
Data: sleep.df
Models:
ran.int: Reaction ~ Days + (1 | Subject)
ran.coef: Reaction ~ Days + (Days | Subject)
         Df    AIC    BIC  logLik  Chisq Chi Df Pr(>Chisq)
ran.int   4 1802.1 1814.9 -897.05
ran.coef  6 1764.0 1783.1 -875.99 42.113      2  7.164e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1
```

Output 2b) Coefficients

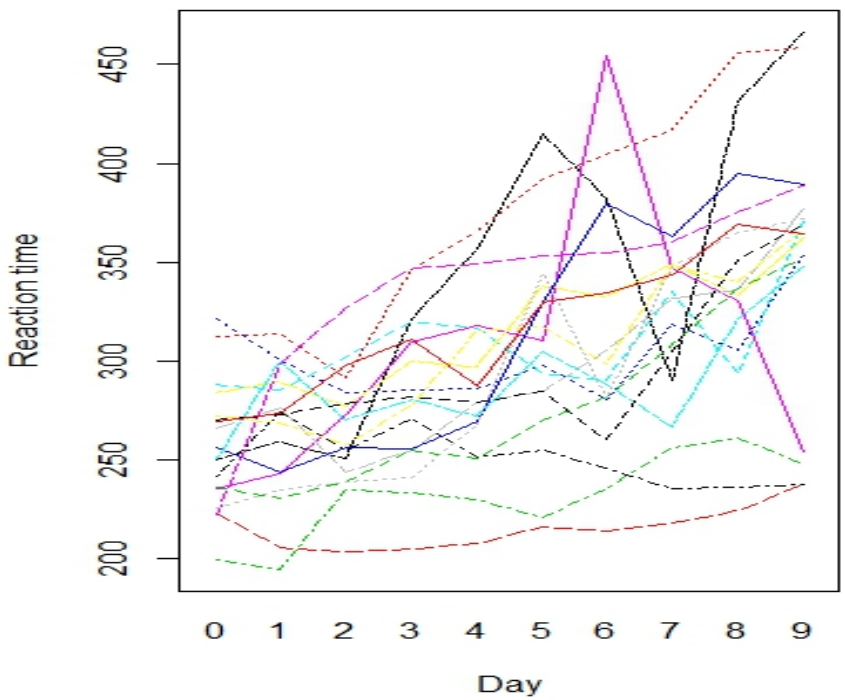
```
> summary(ran.coef)
Linear mixed model fit by REML
Formula: Reaction ~ Days + (Days | Subject)
   Data: sleep.df
  AIC  BIC logLik deviance REMLdev
 1756 1775 -871.8     1752    1744
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.092  24.7405
          Days         35.072   5.9221  0.066
 Residual             654.941  25.5918
Number of obs: 180, groups: Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept)  251.405      6.825   36.84
Days          10.467      1.546    6.77
```

Step 3: Interpreting results

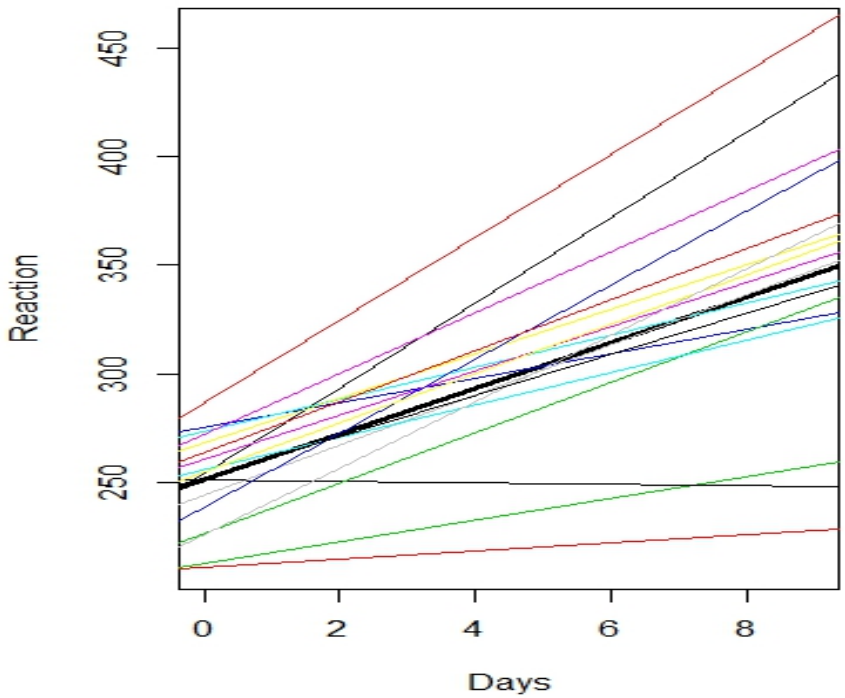
- First we should note that the random coefficients model represents a significant improvement over the random intercept model ($\chi^2_{LR} = 42.113$, $df = 2$, $p < 0.001$)
- BUT is the extra model complexity worth it. Remember AIC is a parsimony measure which accounts for FIT **and** COMPLEXITY
- $AIC_{ranint} = 1802.1$ and $AIC_{rancoef} = 1764$. As random coefficient is lower $\Rightarrow$ a better model (worth it)
- Finally we should note that the fixed effect part of the model is exactly the same ($\beta_0 = 251.4$ and $\beta_{Day} = 10.46$) in both models
- Finally, let's visualize this model to see if it seems a better fit to the data

Random coefficients model: Adequacy

Still not perfect, but looks more realistic.



Real data



Fit (Random coefficients)

Sample size calculation for linear mixed model

- Like other methods, we can calculate sample size for Linear mixed models
- In fact the same formula can be applied to other models used for longitudinal continuous outcomes
- BUT the calculation I am about to show you, only applies when we have a single **Between-subject effect** and a single **Within-subject effect**
- For more complex models we have to go to the rule of thumb methods used for Linear Regression (and General Linear Models)

There is one additional quantity we need in a LMM sample size calculation: **The within-subject correlation, ρ**. This can be estimated from pilot data using the *ICC*, or 'guestimated'

Within-subject correlation

Within-subject correlations is something we go into detail about when we get to Linear Marginal Models, but in the meantime, I will give a rather simple definition:

Within-subject correlation is the extent to which measures taken from a single subject are correlated

For example, if my White blood cell count (compared to you) is high on day 1, it is likely to be high (compared to you) on day 2, etc...

Sample size formula for LMMs

The formula to calculate the sample size for an LMM (one within- and one between-subject effect) is:

$$N_{pergroup} = \frac{2(Z_{\alpha/2} + Z_{\beta})^2(1 + [T-1]\rho)}{T(MCD/\sigma)^2}$$

where:
T is the number of repeated measurements (for $t = 1, 2, \cdots, T$)
ρ is the level of within-subject correlation
$Z_{\alpha/2}, Z_{\beta}, \sigma$ and MCD are defined as before

Sample size for LMMs: Worked example

Let's consider the Autism data. Recall:

- The outcome is level of socialization (VSAE score)
- The **Within-subject** effect is **Age** (2,3,5,9,13yo)
- The **Between-subject** effect is **Sicdegp at 2yo** (language score): Low, Med, High

Note that we have **three** 'between-subjects' groups here (low medium and high), but we are used to thinking about the two group case. BUT all this means is that we have to multiply $N_{pergroup}$ by 3 (rather than 2)

Sample size for LMMs: Worked example

Ingredients

Now let's start collecting our *ingredients*:

1. **Type 1 error:** $\alpha = 0.05$
2. **Type 2 error;** $\beta = 0.1 \Rightarrow$**Power=90%**
3. **Standard deviation:** Pilot data (backed by previous studies) suggest that we would expect to see a standard deviation (for a certian age, for a certain langauge group) of 15units: $\sigma = 15$
4. **Minimal clinical difference:** **CONSULATION WITH EXPERTS** suggests that a difference in socialization score (VSAE) less than 10 would not represent a scientifically important difference: **MCD=10**

Now, these are the quantities we are used to, what about the quantities that are specific to repeated mesaures studies?

Sample size for LMMs: Worked example

Ingredients

Continuing...

5. **T:** In this study we are considering 5 repreated measurements (time points: 2, 3, 5, 9 and 13yo): **T=5**
6. ρ**:** Within-subject correlation. This can be a tricky one (particularly if we don't have pilot data). In the absence of knowledge, I will assume there is a moderate level of within subject correlation: $\rho = 0.6$

For a rough guide (when you don't know)

- $\rho = 0.3 \Rightarrow$ Weak within-subject correlation
- $\rho = 0.6 \Rightarrow$ Moderate within-subject correlation
- $\rho = 0.9 \Rightarrow$ Strong within-subject correlation

Sample size for LMMs: Worked example

Calculation

Now we have all of the ingredients, let's preform the calculation:

$$N_{pergroup} = \frac{2(Z_{0.05/2} + Z_{0.1})^2(1 + [T-1]\rho)}{T(MCD/\sigma)^2}$$

$$N_{pergroup} = \frac{2(1.96 + 1.28)^2(1 + [5-1]0.6)}{5(10/15)^2}$$

$$N_{pergroup} = \frac{21(1 + 2.4)}{5(0.666)^2} = \frac{71.4}{2.222} = 32.13$$

So we need 33 children per group $\Rightarrow$ 3 (language group) * 33 = 99 children (observational units).
Now there are 5 times points so we will have **N=5x99=495**

Sample size for LMMs vs cross-sectional

Just out of interest, let see what we would get if we were doing this study cross sectionally (i.e.$\rho = 0$)

$$N_{pergroup} = \frac{2(Z_{0.05/2} + Z_{0.1})^2(1 + [T-1]\rho)}{T(MCD/\sigma)^2}$$

$$N_{pergroup} = \frac{2(Z_{0.05/2} + Z_{0.1})^2(1 + [T-1]0)}{T(MCD/\sigma)^2}$$

$$N_{pergroup} = \frac{2(Z_{0.05/2} + Z_{0.1})^2(1)}{1(MCD/\sigma)^2}$$

$$N_{pergroup} = \frac{2(Z_{0.05/2} + Z_{0.1})^2}{MCD^2/\sigma^2}$$

$$N_{pergroup} = \frac{2(Z_{0.05/2} + Z_{0.1})^2\sigma^2}{MCD^2}$$

Recognize this formula???

Sample size for LMMs vs cross-sectional

Plugging in ingredients (1-4 only)

$$N_{pergroup} = \frac{2(Z_{0.05/2} + Z_{0.1})^2\sigma^2}{MCD^2}$$

$$N_{pergroup} = \frac{2(10.5)15^2}{10^2} = \frac{21 * 225}{100} = 47.25$$

So for a cross-sectional study (comparing the three language groups) we would need N = 3 * 48 = 144 individuals

In terms of obserational UNITS, which is more efficient?

Where to from here

- There is one more model I want to cover when it comes to modelling continuous longitudinal data: The Linear Marginal Model
- BUT I am going to leave this the next session. This is for two reasons:
  1. Brain burn-out
  2. Linear Marginal models are not used that often.
- So why cover them?? Because understanding Linear Marginal models gives a strong insight into how Generalized Estimating Equations (GEEs) work. GEEs are a generalized equivalent of the Linear Marginal model used for categorical correlated outcomes
- I hope in the next session, that we can practice using these models for longitudinal data (LMMs, GEEs, GLMM)

THANK-YOU

Questions