PSY 5939: Longitudinal Data Analysis

Contents

Linear mixed model

Linear mixed model

Linear mixed model

also known as

- Random coefficient models
- Multilevel models
- Nested models
- Hierarchical linear models
- Random effects models

Different names because developed in different disciplines

- Random coefficient from statistics and biostatistics
- Multilevel models from education

Linear mixed model

Model for **non-independent** observations

- These can be longitudinal or cross-sectional

For example

- Several children from the same family
- Multiple schoolchildren with the same teacher

- Employees who work in teams or workgroups
- **Multiple observations from the same individual over time**

Observations from the same family / class / team / person are more similar to one another than observations from different families / classes / teams / persons

Linear mixed models

Non-independent observations means that there is some **redundancy** (or correlation) between observations

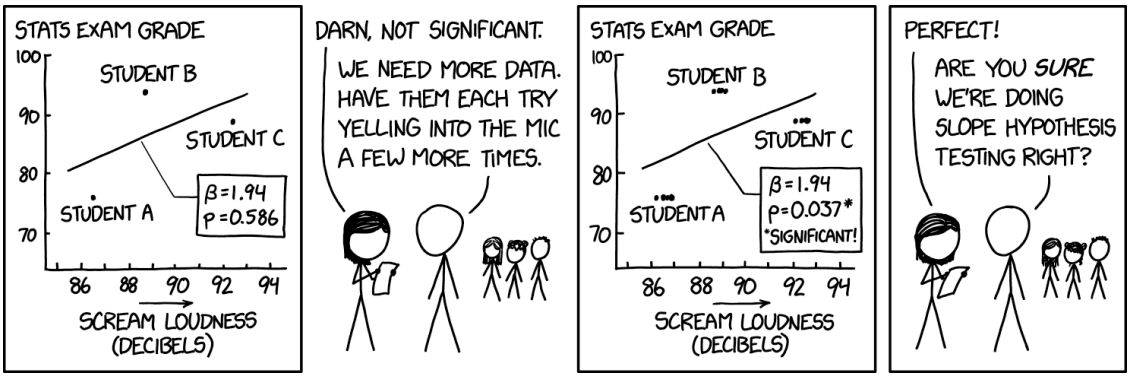
The effective sample size is smaller that the actual sample size

- There are 100 subjects but we only have (for example) 72 subjects' worth of information, due to correlations between observations

The smaller effective sample size means that the standard error is **underestimated** if you ignore non-independence (and use regression)

- How much the standard errors are underestimated depends on how much the observations are related to one another

The wrong way to deal with this



https://xkcd.com/2533/

Linear mixed model: motivation

The linear mixed model (LMM) is an extension of the general linear model (GLM)

Partitions variation, just like ANOVA and regression

But **more ways** to partition and **more control** over the form

How are participants related to one another?

- Between-subjects ANOVA uses "independence"
- Repeated-measures ANOVA uses "compound symmetry" (and sphericity)
- Linear mixed model has **two different approaches** to sorting out the additional variation due to repeated measures

Linear mixed model: equations

$$Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$$

Fixed effects: $\mathbf{X}\beta$

- $\mathbf{X}\beta$ is a linear combination of predictors (matrix $\mathbf{X}$) and regression coefficients (vector β)

Two places where additional variation (i.e., non-independence) is modeled

- **Random effects:** $\mathbf{Z}\gamma$
- **Correlated residuals:** ϵ
- (There are times you might use both, but we're only going to talk about them separately)

Random effects and the G matrix

Random effects: motivation

Linear regression

Linear regression has two important characteristics

1. All observations are assumed to be **independent**
    - Non-independence (e.g., repeated measures) violates this
2. There is a **single effect** of X on Y
    - We're moving into a modeling framework that focuses on **individual effects**
    - There isn't a single effect of X on Y, but each person (in our situation) might have a **different** relationship between X and Y

Individual effects

In linear regression, there is a single, average effect of X on Y

For random effects models, the focus changes to **individual effects**

- Cross sectional: How does a family's / class' / workgroup's outcome (Y) change with some predictor (X)?
- **Longitudinal**: How does *a person's score* (Y) change with time (X)?

We can also talk about average effects, but they are the **average** of individual effects

Conceptually first (pictures), then equations

Conceptual mixed model

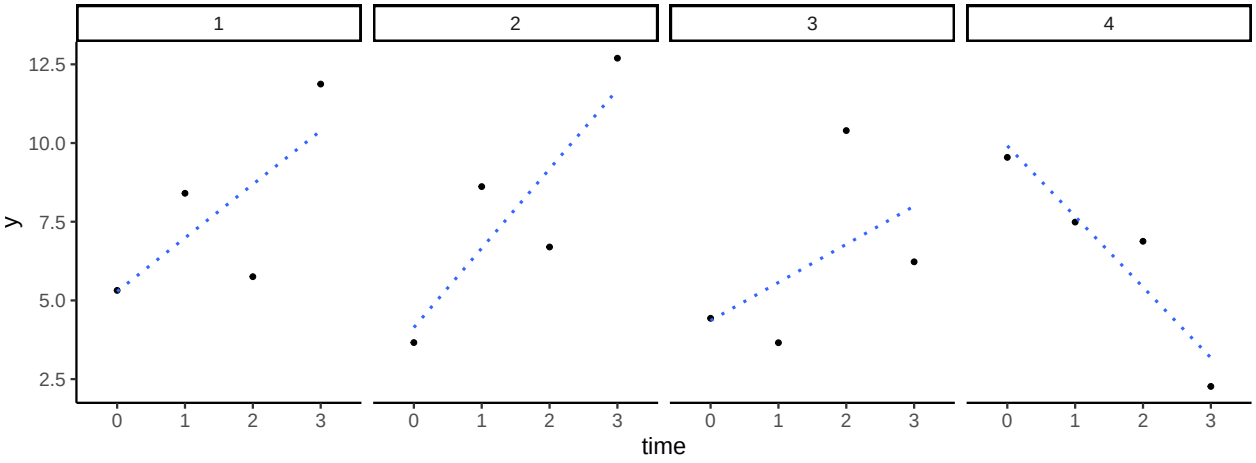
Conceptually

Random effects models for longitudinal data perform a regression on **each** participant's data

- Time is the predictor
- Whatever outcome you're measuring is the outcome

Conceptually, you have a separate regression for each participant in the study

Time versus outcome for each participant



Assumptions

Figures of individual slopes are a good way to think about the model

But "individual regressions" violate the linear regression assumption of non-independence

- For each line, the observations are all from the same person
- So we can't just estimate each person's regression and report each of them

But non-independence is actually only a problem for estimating **standard errors**

- We can use the estimates of **intercepts** and **slopes**

Spaghetti plot

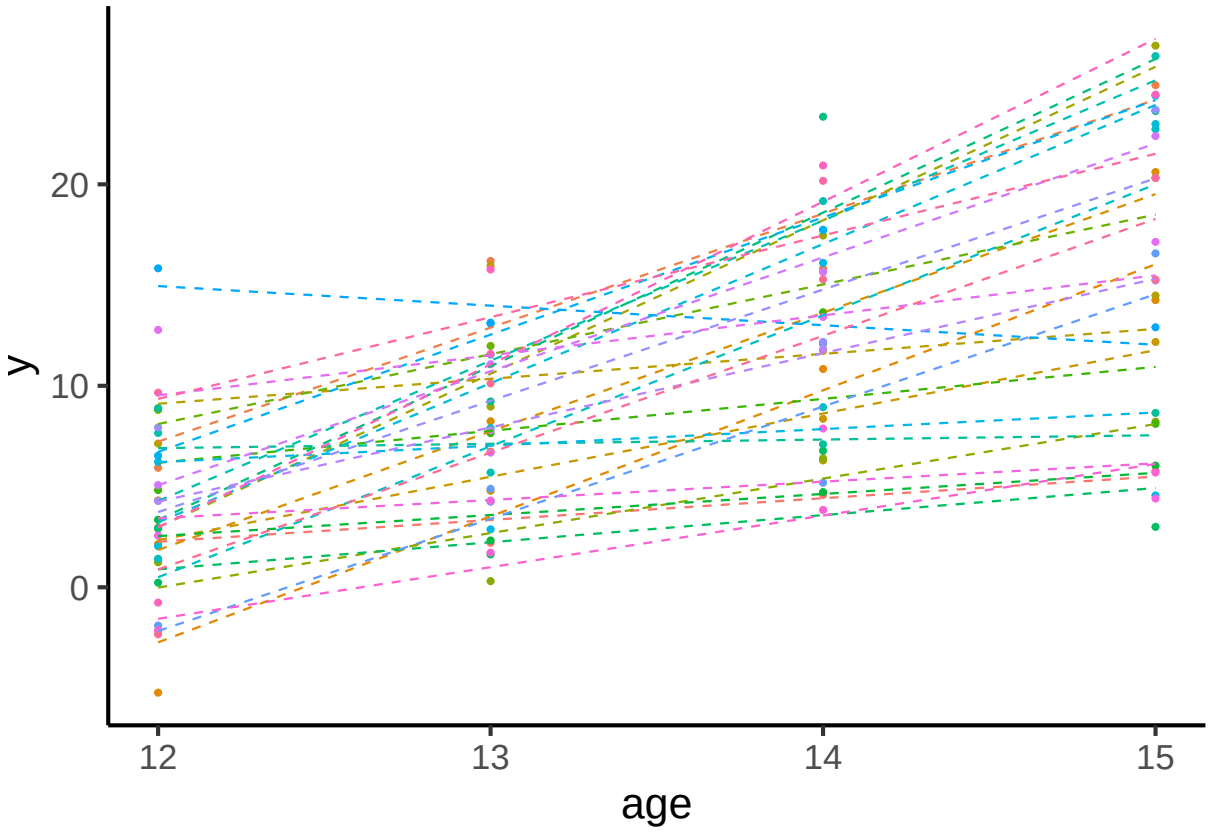
As we move into these models (as well as latent growth models), **looking at your data** will become more important

Looking at the combined individual regressions for all participants can help you build your model

A plot with all the individual regressions is called a **spaghetti plot**

- Estimates of the intercepts
- Estimates of the slopes
- Show you what is going on in your data

Spaghetti plot



Spaghetti plot

Looking at this, we can see that individuals vary

- Mean level / intercept
- Slope (with respect to time)

But we're not going to report individual intercepts and slopes for each person

- We want some summary of the variability information
- We'll estimate **variances** of the slopes and intercepts

Linear mixed model: equations

The linear mixed model: $Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$

1. $\mathbf{X}\beta$ are the **fixed effects**

Think regression:

- $\mathbf{X}$ are predictors
- βs are the coefficients

We'll come back to these more later

Linear mixed model: equations

The linear mixed model: $Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$

2. $\mathbf{Z}\gamma$ are the **random effects** (γ is gamma)

Spaghetti plot: individuals vary. But how?

Z are predictors (of individual variability)

- What do people vary **with respect to**?
- Usually in longitudinal models, just **time**

γ is the random effects

- Variance of intercepts and slopes (also covariance)
- Normally distributed with mean 0 and variance **G**: $\gamma \sim N(0, \mathbf{G})$
- **G** is the variance-covariance matrix of intercepts and slopes

Linear mixed model: equations

The linear mixed model: $Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$

3. ϵ is a residual

- Again, think regression residual

Level 1 and level 2

Level 1 and level 2

One of the names is "multilevel model"

- Models data at several levels

Cross-sectional: Children nested within schools

- Several schools, with children from each school
- Children are "level 1" and schools are "level 2"

Longitudinal: Observations nested within subjects

- Several participants, each measured multiple times
- Measurement occasions are "level 1" and subject is "level 2"

Notation

Multilevel model, hierarchical linear model, random effect model

- These are all names for the same thing, but they were developed in different disciplines
- Have different notations: some use β, some use π, etc.

I use a combined notation that is not found in any specific discipline, it is easy to present in an article

- So not quite the LMM notation
- It also uses only Greek letters that you actually know

Level 1 model

Level 1 = measurement occasion: Model of observations over time

$$Y_{ij} = \pi_{0i} + \pi_{1i}(Time_{ij}) + e_{ij}$$

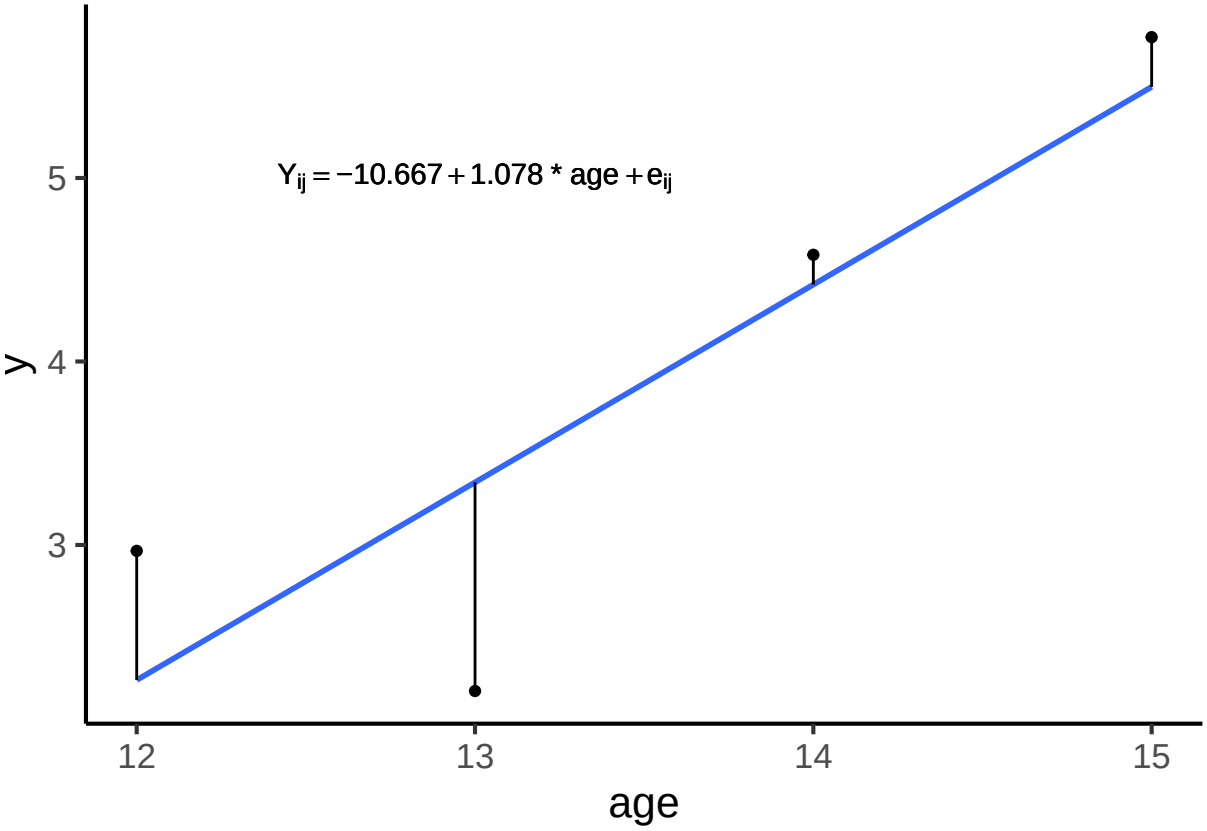
Y_{ij} = outcome score for subject i at occasion j

π_{0i} = intercept for subject i

π_{1i} = slope for subject i

e_{ij} = residual for subject i at occasion j

Level 1 model



Level 2 model

Level 2 = participant: Model of participant level differences

$$\pi_{0i} = \beta_{00} + r_{0i}$$
$$\pi_{1i} = \beta_{10} + r_{1i}$$

π_{0i} = intercept for subject i (from level 1 model)

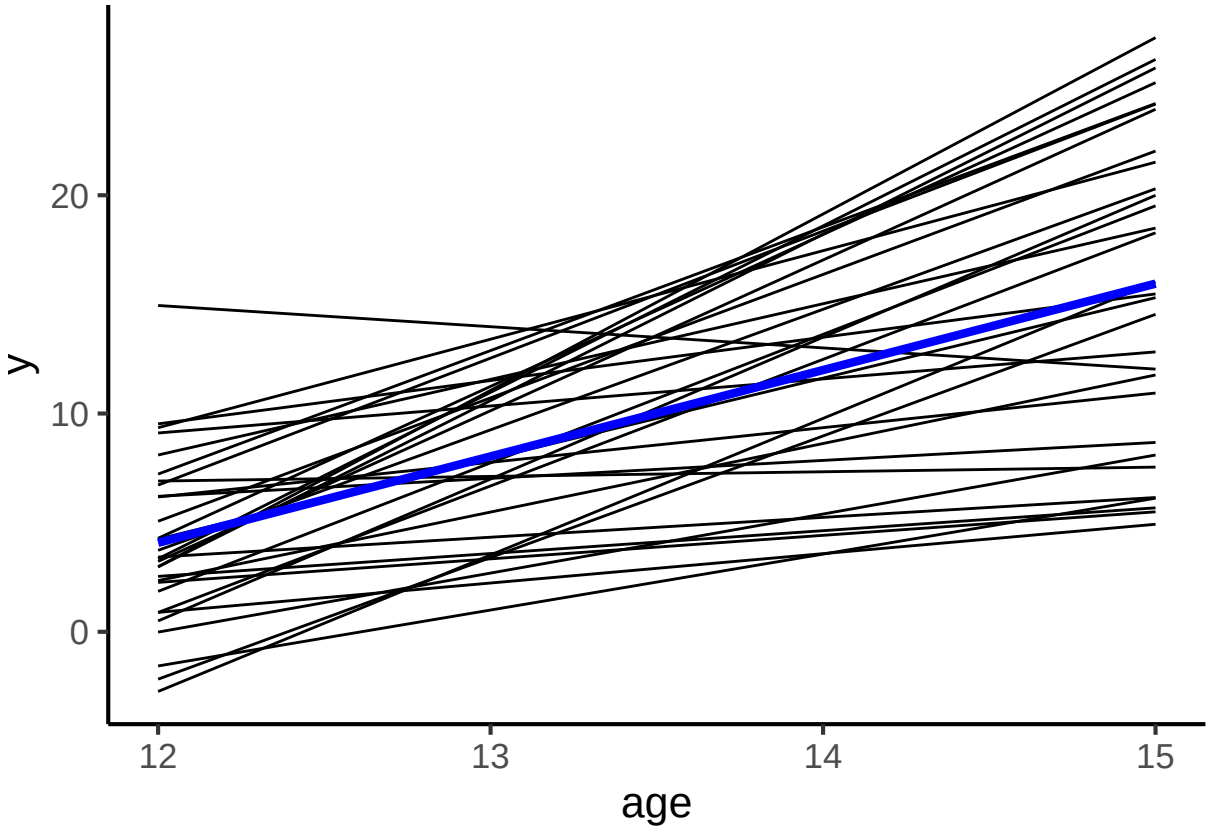
π_{1i} = slope for subject i (from level 1 model)

β_{00} = average intercept, β_{10} = average slope

r_{0i} = individual deviations around the mean intercept

r_{1i} = individual deviations around the mean slope

Level 2 model



Combined model

Substitute level 2 equations into level 1 equation:

$$Y_{ij} = \underline{\pi_{0i}} + \underline{\pi_{1i}}(Time_{ij}) + e_{ij}$$

$$Y_{ij} = \underline{\beta_{00} + r_{0i}} + \underline{\beta_{10} + r_{1i}}(Time_{ij}) + e_{ij}$$

$$Y_{ij} = \beta_{00} + \beta_{10}(Time_{ij}) + r_{0i} + r_{1i}(Time_{ij}) + e_{ij}$$

Everything with a β is a **fixed effect**: these are the $\mathbf{X}\beta$

Everything with a r or e is a **random effect**: these are the $\mathbf{Z}\gamma + \epsilon$

Variance components

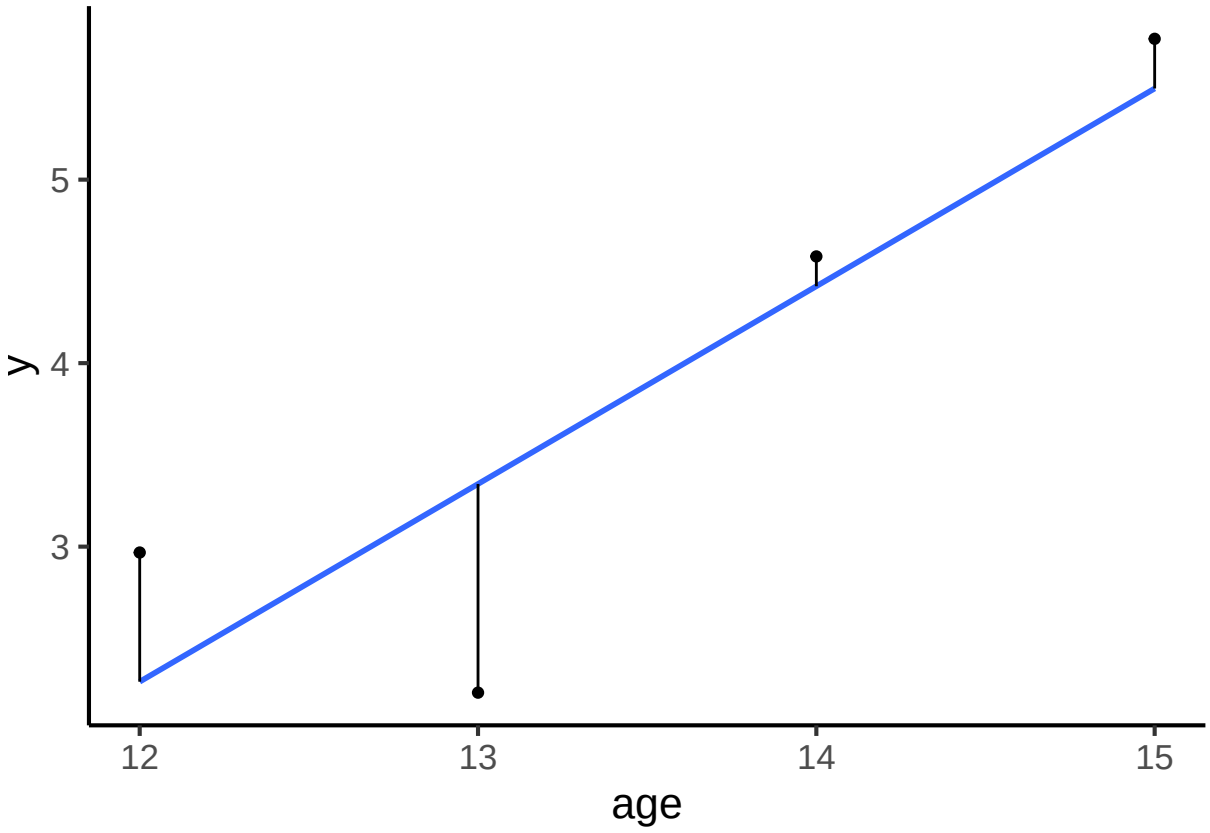
Variance components

There are two levels, and each level has variance components

σ_e^2 is the **level 1 residual variance** (i.e., variance of the residuals for individual trajectories)

- Like the residual variance in linear regression
- Distance between observed data point and individual growth trajectory
- Assumed to be the same for each person, so only the one number

Variance components: level 1



Variance components: level 2

G is the variance-covariance matrix of **individual** intercepts and slopes

- More generally, it's the **level 2 residual variance matrix**

G can take different forms, but here's one we'll use a lot:

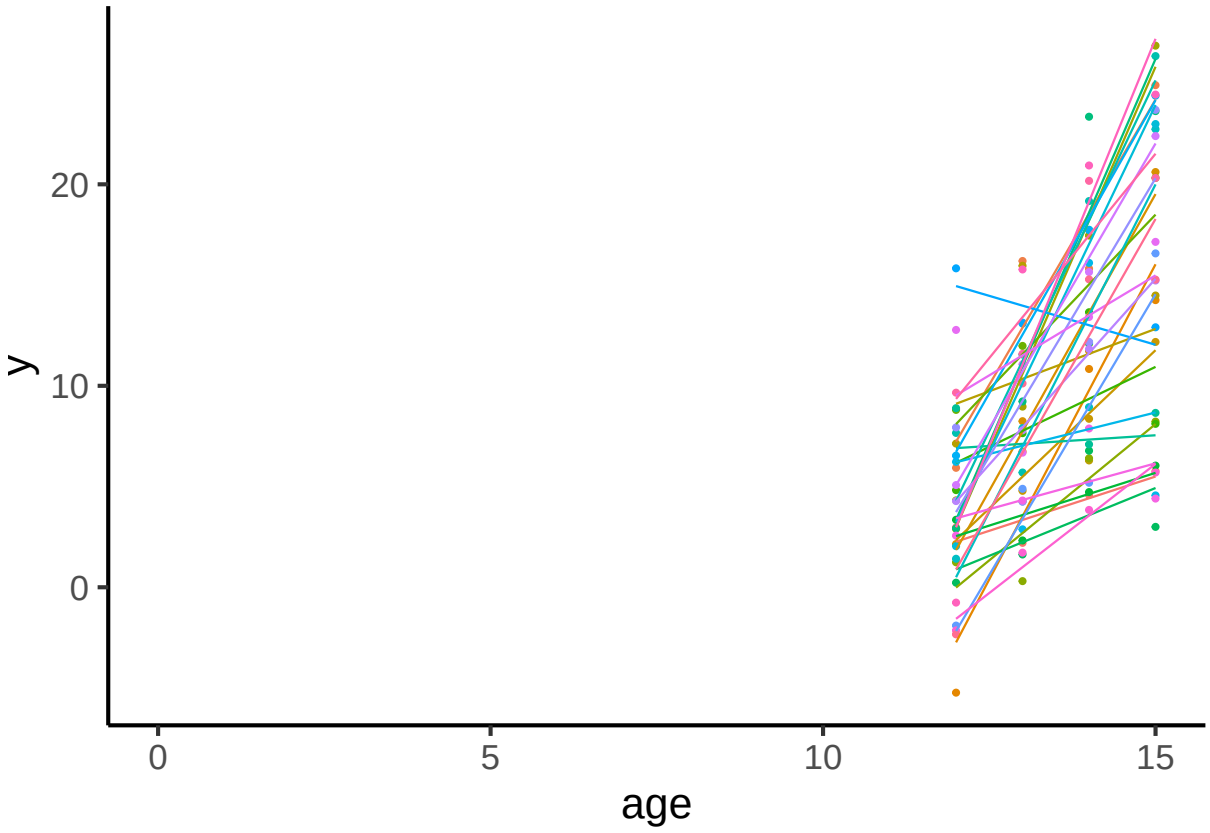
$$\begin{bmatrix} \sigma^2_{r_{0i}} & \\ \sigma_{r_{0i}r_{1i}} & \sigma^2_{r_{1i}} \end{bmatrix}$$

$\sigma^2_{r_{0i}}$ is the variance of the level 2 intercept residuals

$\sigma^2_{r_{1i}}$ is the variance of the level 2 slope residuals

$\sigma_{r_{0i}r_{1i}}$ is the covariance between the level 2 intercept and level 2 slope residuals

Variance components



Summary

Summarizing LMM

Linear mixed model: $Y = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon$

X are the predictors of fixed effects

- Like predictors you're used to
- Includes **time** in longitudinal models

Z are the predictors of random effects

- Usually a subset of variables in **X** (but don't have to be)
- Most commonly **time** in longitudinal models

Longitudinal mixed model:

$Y_{ij} = \beta_{00} + \beta_{10}(Time_{ij}) + r_{0i} + r_{1i}(Time_{ij}) + e_{ij}$

Predictors in **X**: intercept (β_{00}) and time (β_{10})

Predictors in **Z**: intercept (r_{0i}) and time (r_{1i})

Time as a predictor

Time as a predictor

Time in mixed models

In repeated measures ANOVA and some other models

- *Time* is discrete (and often equally spaced)
- Everyone is evaluated together at each wave

Mixed models treat time as **just another (continuous) predictor**

- Time is continuous
- Time does not need to be equally space (and we can model that)
- Potentially different schedule for each individual

Time is the **most important predictor** in longitudinal models

Centering time

Time = 0

Time is a predictor

How time is included as a predictor affects our interpretation of

- Intercept
- Variance of intercept
- Covariance of intercept and slope

Recall from linear regression:

Intercept is the predicted value of Y when $X = 0$

If the *time* variable is age: do you really care about age = 0?

Centering time

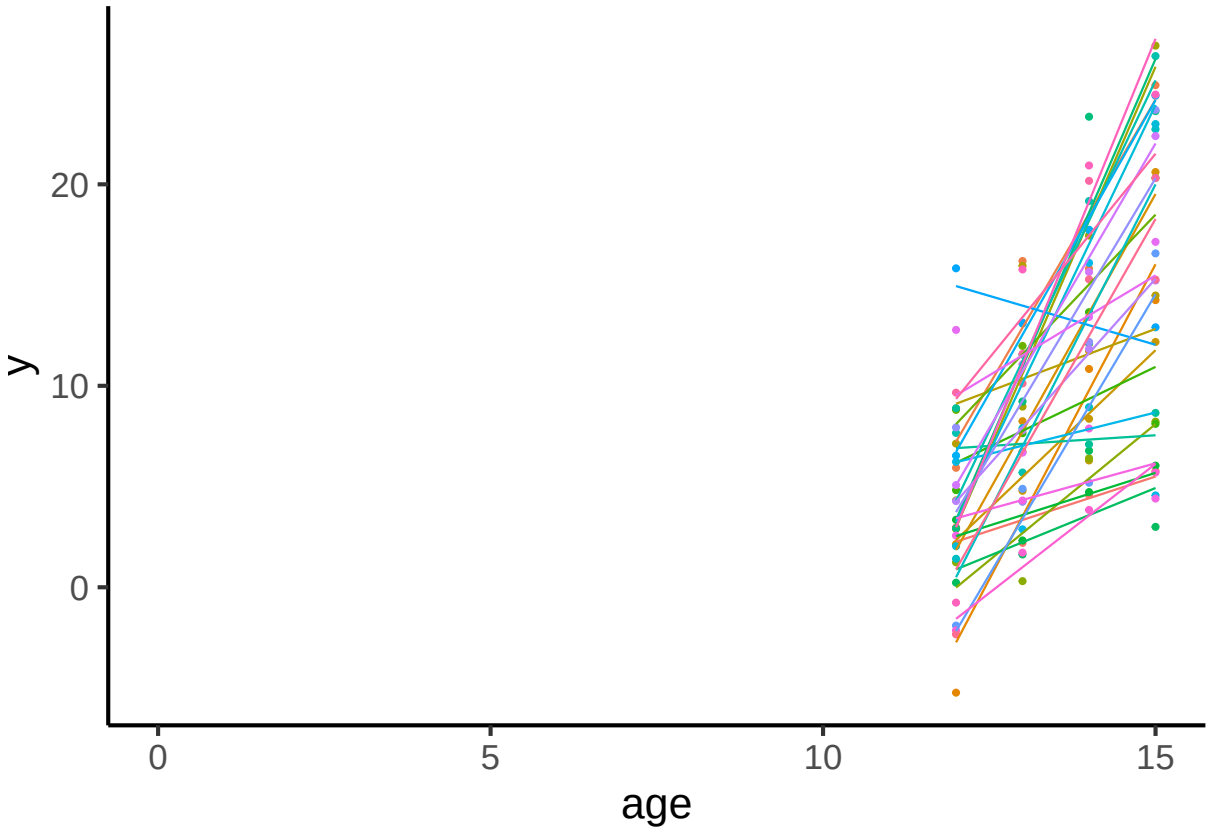
You can (and typically **should**) center the time variable to improve interpretation of your model estimates

Interactions in linear regression normally use "mean centering"

- Centering for longitudinal growth models is more complex
- You will have to decide **where** to center and **justify** that decision

Where you center typically depends on the specific **research question** you're asking

Centering options



Centering example

No centering

- Intercept refers to status at birth (age = 0)

Center at initial status

- Intercept refers to status at age 12

Center at final status

- Intercept refers to status at age 15

Center at some other age

- Some specific age that is substantively relevant
- Intercept refers to status at that age

Trajectory shape

Linear change

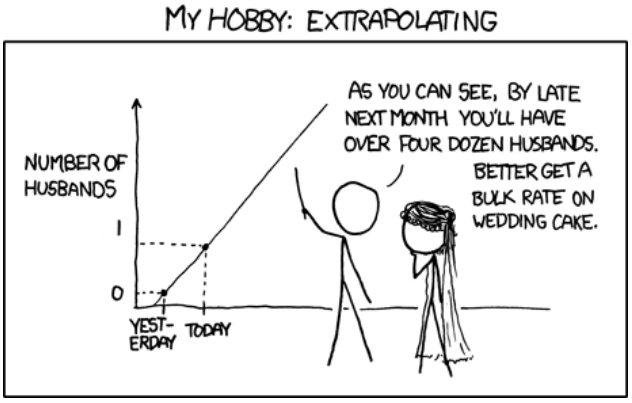
So far, we've been talking about linear change

- It's simple
- Many things change (roughly) linearly
- But that's not all there is

Remember that the trajectory shape is in the level 1 equation:

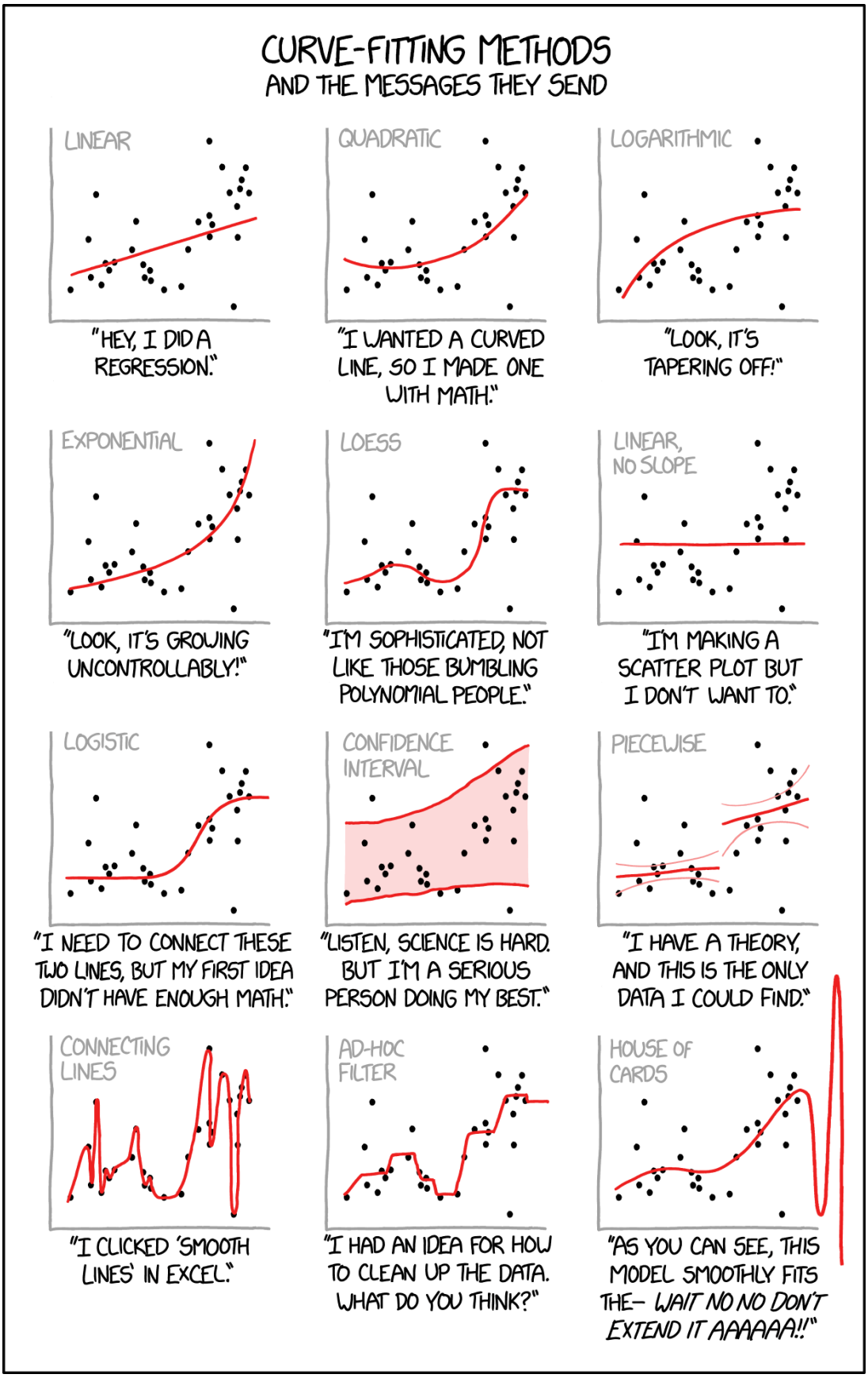
$$Y_{ij} = \pi_{0i} + \pi_{1i}(Time_{ij}) + e_{ij}$$

Is it linear?



http://www.xkcd/605

If not linear, then what?



https://xkcd.com/2048/

Non-linear effects

Non-linear effects of time

Many phases of development or change are non-linear:

- Increase followed by plateau
- Decrease to a set point
- Sometimes reflect floor or ceiling effects

Two main approaches to this: polynomial (quadratic) and ln transform

- Remember that trajectories are modeled in the level 1 equation

Quadratic level 1 equation

$$Y_{ij} = \pi_{0i} + \pi_{1i}(time_{ij}) + \underline{\pi_{2i}(time_{ij})^2} + e_{ij}$$

Linear and quadratic effects

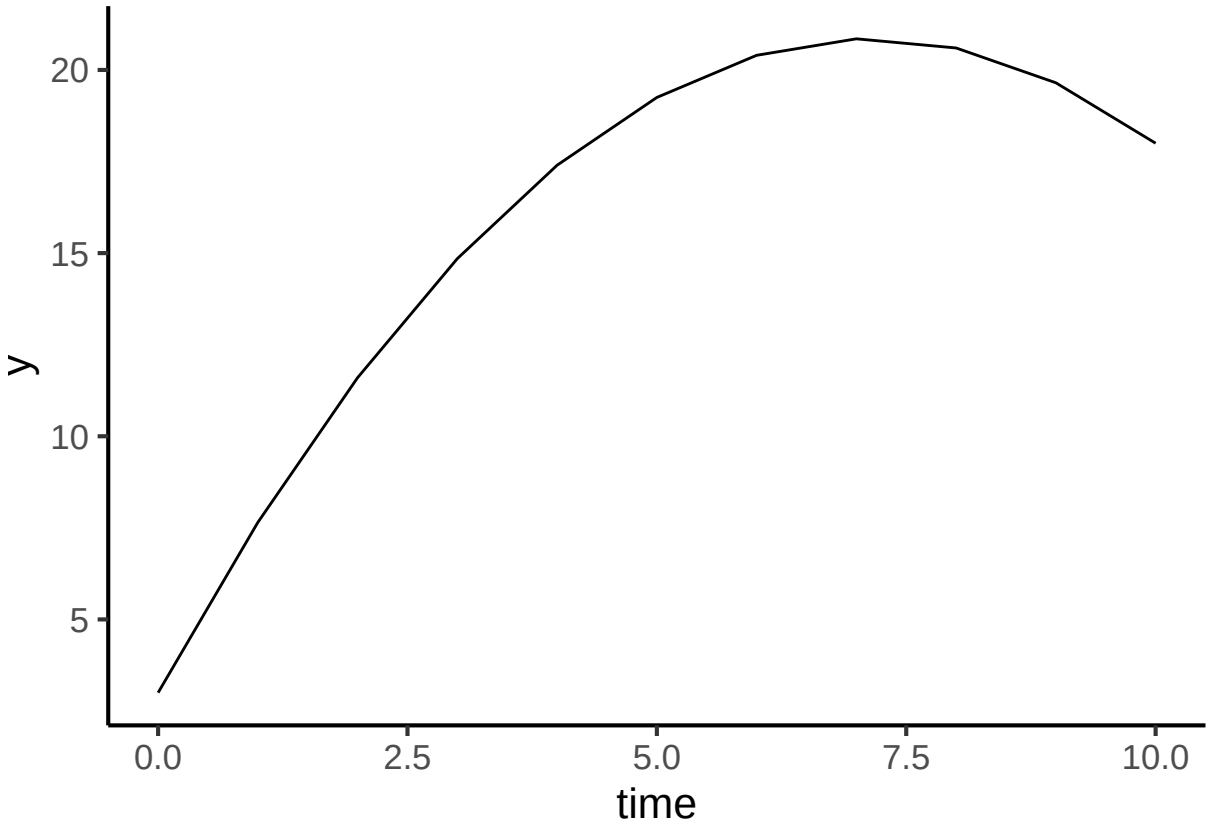
Quadratic trends

- Must be interpreted with the accompanying linear trend
  - This is like interpreting main effects when there's an interaction – they depend on each other
- In general: positive = smiley face and negative = frowny face

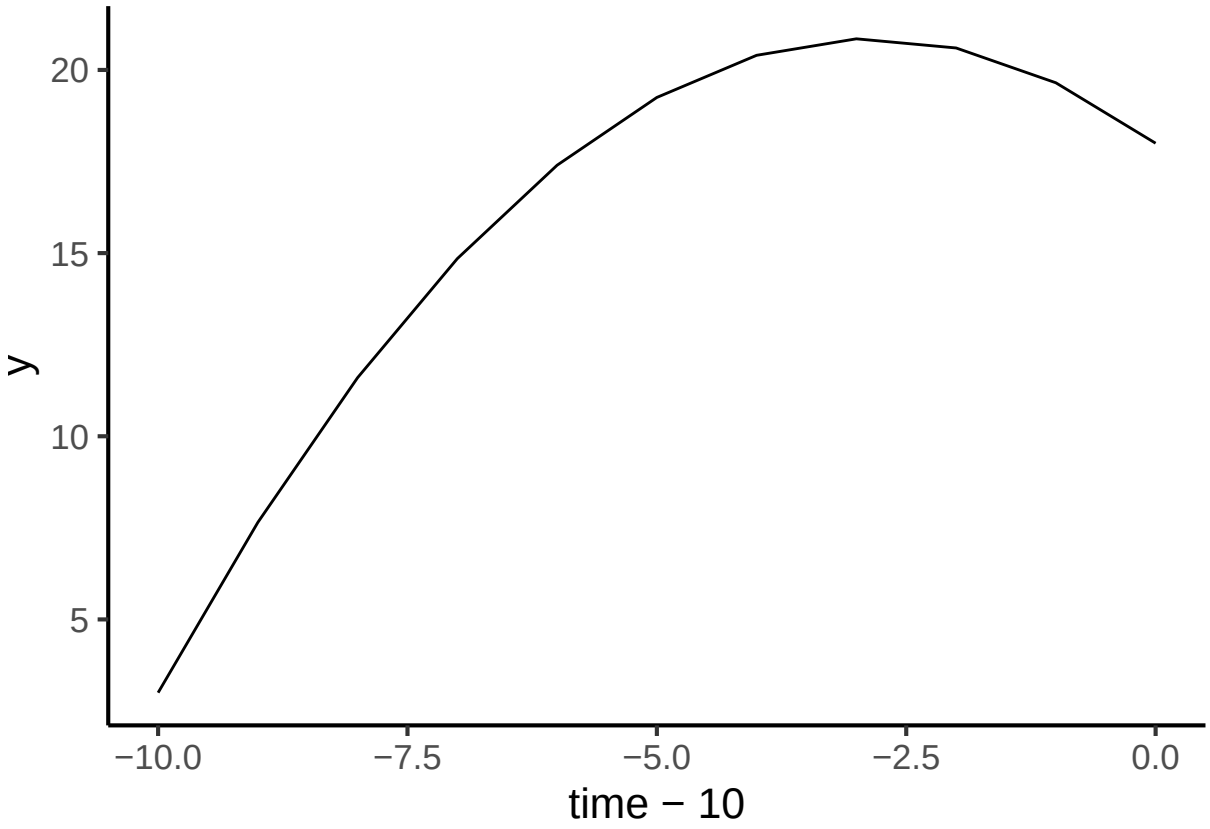
Linear trends

- Positive = increasing
- Negative = decreasing
- Note that the linear effect that accompanies a quadratic effect is the linear effect **at $\mathbf{X = 0}$** (just like with an interaction)

Centering time matters



Centering time matters



Quadratic level 2 equations

People vary in their intercepts (wherever time is centered)

$$\pi_{0i} = \beta_{00} + r_{0i}$$

People vary in their **linear** slopes (wherever time is centered)

$$\pi_{1i} = \beta_{10} + r_{1i}$$

People vary in their quadratic curvature

$$\underline{\pi_{2i} = \beta_{20} + r_{2i}}$$

Quadratic model: variance components

σ_e^2 is the level 1 residual variance (as before)

The level 2 covariance matrix is now 3x3

$$\begin{bmatrix} {\sigma_{r_{0i}}}^2 & & \\ \sigma_{r_{0i}r_{1i}} & {\sigma_{r_{1i}}}^2 & \\ \sigma_{r_{0i}r_{2i}} & \sigma_{r_{1i}r_{2i}} & {\sigma_{r_{2i}}}^2 \end{bmatrix}$$

New elements in the **bottom row**:

$\sigma_{r_{0i}r_{2i}}$ is the covariance between intercept and quadratic slopes

$\sigma_{r_{1i}r_{2i}}$ is the covariance between linear and quadratic slopes

$\sigma_{r_{2i}}{}^2$ is the quadratic slope variance

- How much do people vary in their **curvature**?

Issues in the quadratic model

Often, there is curvature on average

- β_{20} (fixed effect) is significant

but there is not much **variability** in the curvature

- σ^2_{r2i} is nearly 0
- Model fails to converge

Unfortunately, SAS / SPSS / R don't have many options to help

- You can really only **remove** the quadratic variance

Quadratic level 2 equations: remove quadratic variance

People vary in their intercepts (wherever time is centered)

$$\pi_{0i} = \beta_{00} + r_{0i}$$

People vary in their **linear** slopes (wherever time is centered)

$$\pi_{1i} = \beta_{10} + r_{1i}$$

People **don't** vary in their quadratic curvature (no r_{2i} term)

$$\pi_{2i} = \beta_{20}$$

Quadratic level 2 equations: remove quadratic variance

Without the quadratic variance, **G** reduces to

$$\begin{bmatrix} \sigma_{r_{0i}}{}^2 & \\ \sigma_{r_{0i}r_{1i}} & \sigma_{r_{1i}}{}^2 \end{bmatrix}$$

Natural log (ln) transform

Another option for non-linear change is to use $ln(time)$ as predictor

Pros

- Can better reflect increase / decrease followed by plateau because it's not "U" shaped
- Only one slope to worry about
- "Monotonically increasing or decreasing"

Cons

- Somewhat harder to interpret (than linear or quadratic)

- 1 unit increase in $ln(time)$ is an increase in time of about 2.71 units (e)
- Centering is tricky (literally)

ln(time) level 1 equation

$$Y_{ij} = \pi_{0i} + \underline{\pi_{1i}(ln(time_{ij}))} + e_{ij}$$

Centering with ln(time)

$ln(0)$ is undefined

If 0 is in your time variable, it will be undefined

- If your time variable is "time since intervention" then baseline measure is time = 0

How to make this happen?

- Literally trick the program
- Add a small constant to time, then do ln transform
- $ln(0)$ is undefined but $ln(0.00001)$ is fine