The Role of Time in Educational Experiments: Reflections on Longitudinal Experiments

(including Growth Modeling)

Chris Rhoads
University of Connecticut

Presented at IES Cluster Randomized Trial Workshop

Northwestern University

June 2022

What Are Longitudinal Experiments?

- Experiments with repeated measurements of an outcome on the same units.
  - Could be schools, teachers, students, all of the above.
  - We consider here experiments where individuals remain in a single treatment group throughout the study.
    - Additional complexities are introduced if this is not the case.
  - We need to be careful about our design/how we conceptualize treatment to meet this criterion while maintaining an RCT.

Running example

- 3 years of math PD delivered to teachers in grades 3-5 in intervention group.
  - i.e. T is delivered over course of 3 years
  - Participating teachers are those who consented (not all teachers).
- Could be school level OR teacher level randomization
- Assume students don't leave school, teachers don't change grade, etc.
  - i.e. idealized example.
- Numerical examples focus on impact of 1st year of PD on students of 3rd grade teachers in Y1 of implementation.
  - But consider possible complications in interpretation as we move across 3 years of data collection.
    - Construct validity of cause!
- Assume district uses something like iready:
  - Collected 3 times per year so 9 times over course of 3 years.
  - Label Y_1 , … , Y_9 .

A fundamental tension in educational research

- We often don’t know when effects of treatments (if they exist) will reveal themselves, or if those effects will be sustained.
- Nor do we know how much exposure/dose is necessary to have an impact.
  - Especially interventions that are mediated through teachers.
- *IES encourages you to measure education outcomes at multiple timepoints to determine if short-term changes in education outcomes are sustained over time. If it is not possible to do this in the current study design, include activities that may help you apply to IES for an additional Follow-Up grant, such as maintaining contact with schools and study participants.*
  - *From FY 2020 IES Research Grant RFP: Initial Efficacy and Follow-up section.*
    - *Language in current version is worse.*

However..

- System not very stable.
  - Students move/skip grades/repeat grades.
  - Teachers:
    - Change grades
    - Change jobs
    - Leave profession
- Superintendents/principals/school board members change.
- If you use administrative data, systems getting better at tracking students.
- Not so much for teachers.

Difficulties in longitudinal experiments

- Very hard to maintain implementation and preserve randomization as time goes on.
- Best case scenario is probably…

Example: TN class size (STAR)

Approaching the ideal…

1. Students randomized to small class, large class or large class with aide.
2. Students remain in same condition K-3rd grade.
3. Students randomized to classrooms/teachers.
    - Teachers randomized to condition each year.
    - Students re-randomized to teachers each year.
    - Crucial, otherwise teacher effects confounded with treatment effects.
4. Funded by legislature, district-wide for participating districts.
    - No worries about superintendent/principal buy-in.
    - Student mobility not-problematic.
5. Not mediated through teachers (e.g. not teacher PD)
    - Teacher attrition not-problematic.

Fundamental tension (for all non-STAR researchers)

- I don’t have solution…
- Many interesting questions about longitudinal experiments:
  - How best to design them?
    - E.g. level of randomization.
  - Is the additional information worth the internal validity cost due to attrition?
  - Is it more important to follow schools or teachers or students over time?
    - “Schools” is far and away the easiest but maybe the least interesting.

Reasons to collect longitudinal data in experiments

1. More than one discrete endpoint is of interest. Questions include:
   a) How long does it take for effect of intervention to "take hold"? (T tries to change teacher behavior).
   b) If there is an initial effect of the intervention, does it "fade out" over time? (early childhood interventions).

<u>Examples</u>:

i) Experiments with immediate and delayed posttests.

ii) Experiments that track individuals over many performance periods.

- Eg. TN class size.
  - Kids have been followed up to adulthood. Chetty's work.

Reasons to collect longitudinal data in experiments

2. Increase the precision (reliability) of measurement.

Example: Direct observations of teacher behavior.

Why are repeated measurements necessary?

- Too much noise in a single measurement.
  - Observer coding varies (inter-rater reliability <1)
  - Behavior itself varies.
    - MET study (Gates):
      - Not only are multiple observers needed.
      - Teachers need to be observed multiple times.

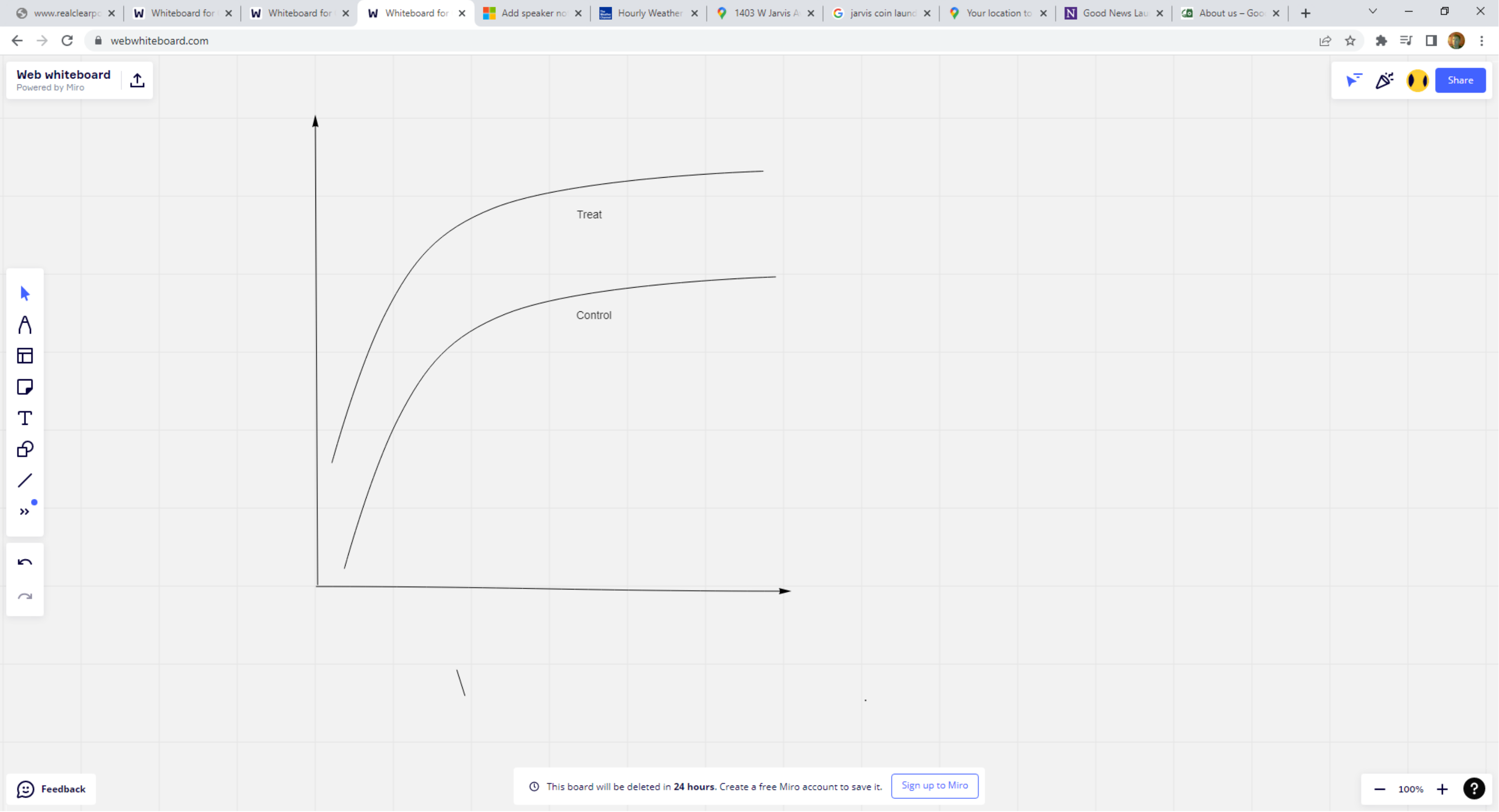
Reasons to collect longitudinal data in experiments

3. The time course of treatment effect (growth trajectory) is of interest (e. g., an intervention is intended to increase the rate of vocabulary acquisition in preschool children).
   - That is, the intervention explicitly aims to impact growth rate.
   - Everyone will eventually learn the words, we want to change how fast kids learn them.
   - (Differential) Curvature in growth function may contain important information about when intervention “takes hold”.

Example

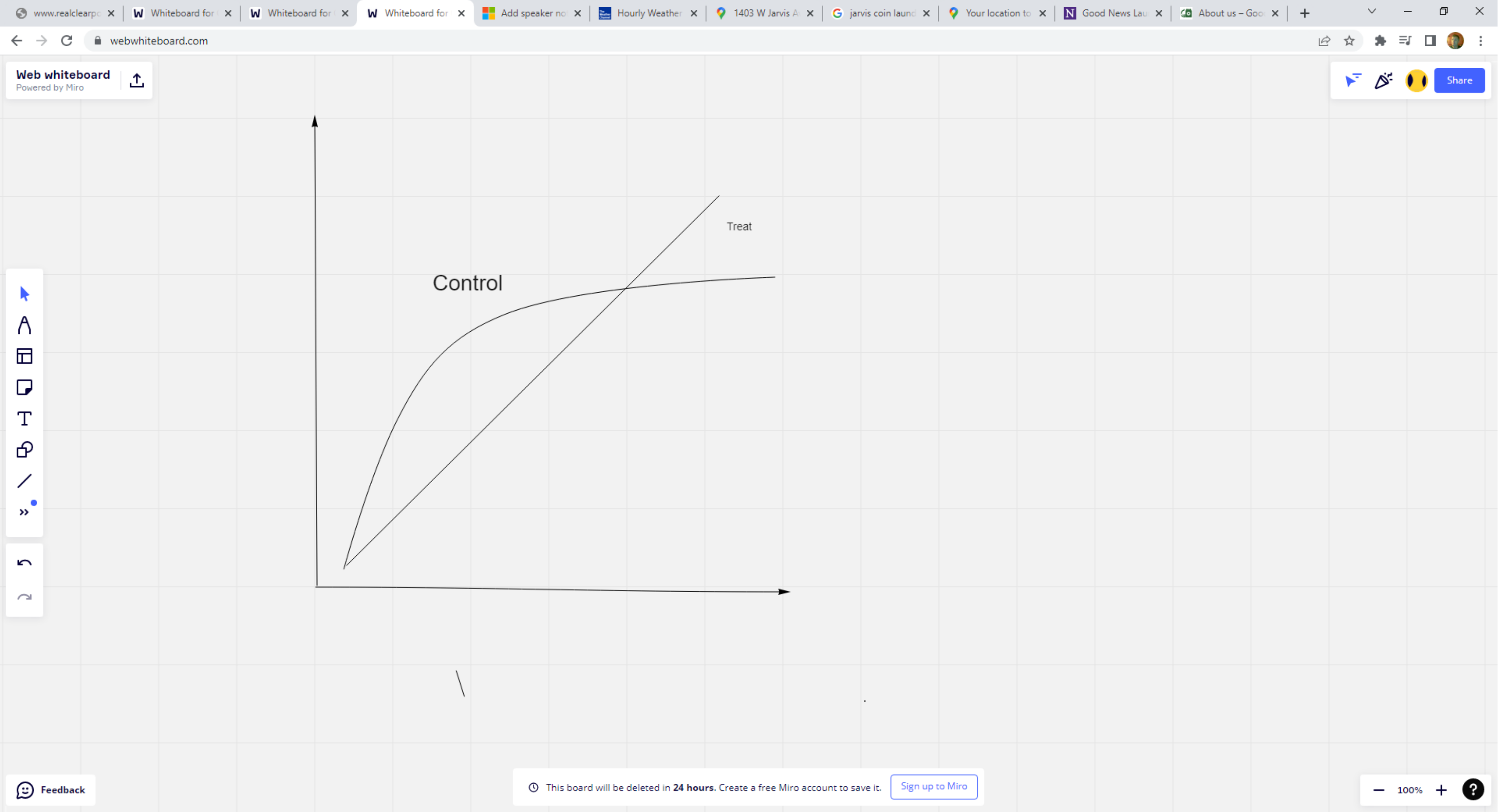
- Spse BAU curve is quadratic.
- We want to know:
  - Does treatment simply shift curve?
    - i.e. Same effect on slope and curvature.
  - Or does it flatten curve?
    - Possibly resulting in negative treatment effects early and positive one later.

Case 1: shift



www.realclearp
Whiteboard for
Whiteboard for
Whiteboard for
Add speaker no
Hourly Weather
1403 W Jarvis A
jarvis coin launc
Your location to
Good News Lau
About us – Goo
webwhiteboard.com
Web whiteboard
Powered by Miro
Share
Treat
Control
Feedback
This board will be deleted in 24 hours. Create a free Miro account to save it.
Sign up to Miro
100%

Case 2: flatten



This board will be deleted in **24 hours**. Create a free Miro account to save it.

Summary

- Three distinct reasons one might consider collecting longitudinal data within context of experiments.

1. Status of outcome at different times is of interest.
2. General measure of status desired but no one measure sufficiently dependable.
3. Want to look at growth trajectories over time.

The ability to answer additional questions (about growth, fade-out, latency of effects) or to improve reliability are the “pros” of longitudinal measurement.

What are the cons?

Reasons to hesitate

1. It will cost more money to follow people for a longer period of time/take more measurement. Is the benefit worth the expense?
   - Do I get enough additional knowledge from the data collected at the additional time points?

- In one sense more data is always better.
- However, trying to collect lots of data points may restrict our ability to do other things, eg.
  - Recruit more schools.
  - Spend some time validating fidelity measures.
  - Etc.

Reasons to hesitate

2. Attrition

In educational studies tracking students and teachers across years can be particularly problematic.

- Espec. middle school years for students.
  - Transition to next school non-standard (8th, 9th, etc.)
- Espec. Elementary for teachers.
  - Teachers often change grade level.
- Might need to consider (sometimes expensive) incentives to keep folks in studies.

Consequences of attrition

- In cross sectional study it is undesirable.
- In longitudinal study it is <u>really</u> undesirable.
  - Throwing out a case b/c student dropped out B4 fourth year means you lose 3 years of data
    - you may not necessarily do this (e.g. growth modeling approaches can handle unbalanced data).
      - However, bias may be creeping into your carefully designed experiment.
    - It may be worth trying to track the folks who move in order to avoid this. <u>But</u> this will get expensive.

Reasons to hesitate

3. Opportunities for circumstances to intervene to potentially undermine randomization.
    a) Parents may lobby to get children into a particular group in the second year of a study.
    b) Administrators may intervene to place kids in certain classrooms (C or T) in the second year of a study.

*Of course, these may be more or less of an issue depending on your design and objectives.

Important questions

1. Do you want teachers to have multiple years to implement (to get familiar with treatment)?
    - May need longitudinal models for teacher outcomes but not student outcomes.
        - Repeated cross-sectional design for student outcomes. Handle with fixed effect for time.
2. Do you want to follow the same students over time?
    - You need longitudinal models.
    - ITT models (only ones obviously valid)- code treatment exposure based on initially randomized condition.
3. Do you need students to have multiple years of exposure to same condition?
    - You need to figure out how to make this happen and still preserve randomization.

Tough (related) questions

1. How much exposure to the “special sauce” is necessary (for teachers/students) in order to see effects?
   – Can we maintain treatment-control contrast long enough?
2. How long after initial/sustained exposure until effect shows up on measures?
3. If effects are evident by time=*t,* will they be sustained until time=*t+x*?

Assume longitudinal growth of students is what you care about.

Possible Solutions to tough questions.

1. Convince yourself one year of student exposure to treatment is sufficient AND
   – You aren’t worried about T and C students exposed to students/teachers in opposite condition during longitudinal follow up.
2. Get schools to let you randomize students to classrooms in subsequent years in order to keep in same condition.
   – Good luck with that!

Possible Solutions to #3

(less satisfying, doesn’t really answer Q of itnerest)

3. Redefine causal effects being estimated.
   - Approach outlined in WWC standards for RCTs with cluster level assignment.
   - If risk of bias due to individuals entering or exiting clusters <u>after</u> random assignment, then:
   - We can still get unbiased estimate of randomization effects on <u>clusters</u> (but can’t attribute effects to changes in individuals, could be due to compositional changes in clusters).
   - At best meets standards “with reservations”.

Multilevel Models for longitudinal data

Modeling longitudinal data

- We can view longitudinal data as a type of *multi-level* model.
- Hierarchical/multi-level models:

a) Can solve the problem of students nested within classrooms and schools.

b) Can also help solve problem of measurements nested within individuals.

  – Particularly useful with unbalanced measurement occasions

- Hence will use HLM notation in our discussion of longitudinal experiments

It is OK to keep it simple

Unless different outcomes (measured on the same individuals) are being compared, you don't have to use longitudinal methods/multilevel models!

- You could measure outcomes at different times and look at them one at a time.

But, if different outcomes (measured on the same individuals) are being compared, outcomes are not independent.

This dependence must be accounted for

(for instance, by using a multi-level model with a "measures" level)

Recall

- Three distinct reasons one might consider collecting longitudinal data.

1. Status of outcome at different times is of interest.
   - “Discrete endpoints”
2. General measure of status desired but no one measure sufficiently dependable.
   - “Average several measures”
3. Want to look at growth trajectories over time.

Running example

- Suppose we want to see if treatment effect at end of 3rd year (Y_9) is same as treatment effect at end of 1st year (Y_3).
  - Or we could compare Y_6 and Y_3.

Discrete Endpoints, Schools Assigned (Comparing Early and Delayed Outcome)

Level 1 (measure level)

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} D_{ijk} + \varepsilon_{ijk} \qquad \varepsilon \sim \mathrm{N}(0, \sigma_W^{\,2})$$

Level 2 (individual level)

$$\beta_{0jk} = \gamma_{00k} + \eta_{0jk}$$

$$\boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_I)$$

$$\beta_{1jk} = \gamma_{10k} + \eta_{1jk}$$

Level 3 (school level)

$$\gamma_{00k} = \pi_{000} + \pi_{001} T_k + \xi_{00k}$$

$$\boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_S)$$

$$\gamma_{10k} = \pi_{100} + \pi_{101} T_k + \xi_{10k}$$

Note that the η_{0jk}'s and η_{1jk}'s can be correlated as can the ξ_{00k}'s and ξ_{10k}'s

Most interpretable to code D_{ijk} as +0.5 and -0.5. Then β_{0jk} is individual level avg. of measures and β_{1jk} is individual level difference between early and delayed measure.

Discrete Endpoints, Schools Assigned (Testing for Fade-out/increased effect)

Level 1 (measure level)

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} D_{ijk} + \varepsilon_{ijk}$$

Level 2 (individual level)

$$\beta_{0jk} = \gamma_{00k} + \eta_{0jk}$$

$$\beta_{1jk} = \gamma_{10k} + \eta_{1jk}$$

Level 3 (school level)

$$\gamma_{00k} = \pi_{000} + \pi_{001} T_k + \xi_{00k}$$

$$\gamma_{10k} = \pi_{100} + \pi_{101} T_k + \xi_{10k}$$

Π_{001}= Avg TE across both measurements

Π_{101}= AVG Difference b/tw TE at first and second measurements (i.e fade-out or increased effect)

Variation in avg. of measures across schools

Variation in <u>difference</u> between measures across schools

Discrete Endpoints (Comparing Early and Delayed Outcome)

Note that, in this model, the ε_{ijk}'s *can* be interpreted as measurement errors.

Thus, the η_{0jk}'s are between individual differences in the "true" early scores (an analogous statement is true for the delayed scores). So, the intraclass correlation

$\rho_I = \sigma_I^2/(\sigma_I^2 + \sigma_W^2)$ is a (individual level) reliability coefficient (in measurement sense of reliability).

Then the ξ_{00k}'s are between-school differences on the average true score quantities

the intraclass correlation

$\rho_S = \sigma_S^2/(\sigma_s^2 + \sigma_I^2 + \sigma_W^2)$ can be thought of as a (school level) reliability coefficient.

NB: These interpretations assume homogeneous measurement error across early and delayed time points.

Discrete Endpoints, Schools Assigned.

TEs at distinct times of interest (but you want to account for correlation between measures) Suppress intercept.

Level 1 (measure level)

$$Y_{ijk} = \beta_{0jk} E_{ijk} + \beta_{1jk} D_{ijk} + \varepsilon_{ijk}$$

Level 2 (individual level)

$$\beta_{0jk} = \gamma_{00k} + \eta_{0jk}$$

$$\beta_{1jk} = \gamma_{10k} + \eta_{1jk}$$

Level 3 (school level)

$$\gamma_{00k} = \pi_{000} + \pi_{001} T_k + \xi_{00k}$$

$$\gamma_{10k} = \pi_{100} + \pi_{101} T_k + \xi_{10k}$$

Π_{001}= Avg TE at first measurement

Π_{101}= Avg TE at second measurement

Variation in first measure avg. across schools

Variation in second measure avg across schools

Discrete Endpoints, Individuals Assigned (Comparing Early and Delayed Outcome)

Level 1 (measure level)

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} D_{ijk} + \varepsilon_{ijk} \qquad \varepsilon \sim \mathrm{N}(0, \sigma_W^{2})$$

Level 2 (individual level)

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k} T_j + \eta_{0jk}$$

$$\boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma_I})$$

$$\beta_{1jk} = \gamma_{10k} + \gamma_{11k} T_j + \eta_{1jk}$$

Level 3 (school level)

$$\gamma_{00k} = \pi_{000} + \xi_{00k}$$
$$\gamma_{01k} = \pi_{010} + \xi_{01k} \qquad \boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma_S})$$
$$\gamma_{10k} = \pi_{100} + \xi_{10k}$$
$$\gamma_{11k} = \pi_{110} + \xi_{11k}$$

Note that the η_{0jk}'s and η_{1jk}'s can be correlated as can the ξ_{00k}'s, ξ_{00k}'s, ξ_{10k}'s, and ξ_{11k}'s

Discrete Endpoints, Individuals Assigned (Comparing Early and Delayed Outcome)

Level 1 (measure level)

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} D_{ijk} + \varepsilon_{ijk}$$

Level 2 (individual level)

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k} T_j + \eta_{0jk}$$

$$\beta_{1jk} = \gamma_{10k} + \gamma_{11k} T_j + \eta_{1jk}$$

Level 3 (school level)

$$\gamma_{00k} = \pi_{000} + \xi_{00k}$$
$$\gamma_{01k} = \pi_{010} + \xi_{01k}$$
$$\gamma_{10k} = \pi_{100} + \xi_{10k}$$
$$\gamma_{11k} = \pi_{110} + \xi_{11k}$$

Π_{010}= Avg TE across both measurements

Variation in avg. TE across both measures and across schools

Π_{110}= Difference b/tw avg TE at first and second measurements (i.e fade-out or increased effect)

Variation in difference b/tw avg TE at first and second measurements across schools

- If additional discrete endpoints are of interest, one can add additional dummy variables at level 1.
  - However, I'd prefer just using separate models.
- By being creative with coding can look at whatever measurement contrasts are of interest.

Recall

- Three distinct reasons one might consider collecting longitudinal data.

1. Status of outcome at different times is of interest.
   - “Discrete endpoints”
2. General measure of status desired but no one measure sufficiently dependable.
   - “Average several measures”
3. Want to look at growth trajectories over time.

2. Average of Several Measures

Continue to assume three levels

Measures are nested (clustered) within individuals, individuals are nested (clustered) within schools

Level 1 (measures within individuals)

Level 2 (individuals within schools)

Level 3 (schools)

Let Y_{ijk}, the observation on the ith measure for the jth person in the kth school with p measures per individual

NB: Note, covariates at first level wouldn't really make sense. Why?

Average of Several Measures (Treatment Assigned at the School Level)

Level 1 (measure level)

$$Y_{ijk} = \beta_{0jk} + \varepsilon_{ijk} \qquad \varepsilon \sim \mathrm{N}(0,\ \sigma_W{}^2)$$

Level 2 (individual level)

$$\beta_{0jk} = \gamma_{00k} + \eta_{0jk} \qquad \eta \sim \mathrm{N}(0,\ \sigma_I{}^2)$$

Level 3 (school level)

$$\gamma_{00k} = \pi_{000} + \pi_{001}T_k + \xi_{00k} \qquad \xi \sim \mathrm{N}(0,\ \sigma_S{}^2)$$

Note that π_{001} *is the treatment effect*

Assumptions in above (and subsequent) models

- No change in “true” score across measurement occasions

OR

- If there is a change in true scores, it is OK to average across these changes to create a composite outcome and this “average” is what we want to measure (i.e. it is sufficiently interpretable).

Average of Several Measures (Treatment Assigned at the Individual Level)

Level 1 (measure level)

$$Y_{ijk} = \beta_{0jk} + \varepsilon_{ijk} \qquad \varepsilon \sim \mathrm{N}(0,\ \sigma_W{}^2)$$

Level 2 (individual level)

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k}T_{jk} + \eta_{0jk} \qquad \eta \sim \mathrm{N}(0,\ \sigma_I{}^2)$$

Level 3 (school level)

$$\gamma_{00k} = \pi_{000} + \xi_{00k} \qquad \boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0},\ \boldsymbol{\Sigma_S})$$

$$\gamma_{01k} = \pi_{010} + \xi_{01k}$$

Note that π_{010} *is the treatment effect (*and it may vary across schools).

Average of Several Measures

Note that, in this model, the ε_{ijk}'s *can* be interpreted as like (item level) measurement errors

Then the β_{0jk}'s can be interpreted as individual level “true” scores (for the j^{th} person in the k^{th} school)

Thus the η_{0jk}'s are between individual differences in these “true” scores and the quantity $\rho_I = \sigma_I^2/(\sigma_I^2 + \sigma_W^2/p)$ is a (individual level) reliability coefficient

Then the ξ_{00k}'s are between-school differences on these quantities and the quantity $\rho_S = \sigma_S^2/(\sigma_s^2 + \sigma_I^2 + \sigma_W^2/p)$ is a true (school level) reliability coefficient

Notice: There are *p* measures at level 1, so the level 1 variance component is divided by *p* when computing reliability.

Running example

- We could fit a model with all 9 measurements nested within students.
  - Might be worried about interpretability of this (averaging across three years where growth may have occurred).
- Could combine idea with idea of previous section (discrete endpoints).
  - Ie. Average within year but not across year
  - Then you'd use the models from previous section (with dummies)
    - Two dummies if you want all three years.
    - I.e. the model on next slide

Discrete Endpoints, Schools Assigned (Comparing Y1 to Y2 to Y3 outcomes, 3/year)

Level 1 (measure level)

$$Y_{ijk} = \beta_{0jk} + \beta_{Djk}D_{ijk} + \beta_{Ejk}E_{ijk\,+}\,\varepsilon_{ijk} \qquad \varepsilon \sim \mathrm{N}(0,\, \sigma_W{}^2)$$

Level 2 (individual level)

$$\beta_{0jk} = \gamma_{00k} + \eta_{0jk}$$

$$\boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0},\, \boldsymbol{\Sigma_I})$$

$$\beta_{Djk} = \gamma_{D0k} + \eta_{Djk}$$

$$\beta_{Ejk} = \gamma_{E0k} + \eta_{Ejk}$$

Level 3 (school level)

$$\gamma_{00k} = \pi_{000} + \pi_{001}T_k + \xi_{00k}$$

$$\boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0},\, \boldsymbol{\Sigma_S})$$

$$\gamma_{D0k} = \pi_{D00} + \pi_{D01}T_k + \xi_{D0k}$$

$$\gamma_{E0k} = \pi_{E00} + \pi_{E01}T_k + \xi_{E0k}$$

Side note

- One could simply average up all the measures and ignore the measure level altogether.
- This is perfectly acceptable.
- The virtue of the hierarchical modeling approach is mainly:
  1. the ability to get reliability coefficients from the variance components.
  2. can easily handle case where not everyone has same # of measurements.

Recall

- Three distinct reasons one might consider collecting longitudinal data.

1. Status of outcome at different times is of interest.
   - "Discrete endpoints"
2. General measure of status desired but no one measure sufficiently dependable.
   - "Average several measures"
3. Want to look at growth trajectories over time.

Before you gather data for growth modeling

1. Be sure the outcome variable that you will measure repeatedly can be compared across time (vertical scaling).
   - Can also be important for comparing early and delayed outcomes, depending on time lag.

- Example: 1st grade vocabulary test.
  - Given to 1st graders it may measure “vocabulary knowledge”.
  - Given to 5th graders it may measure “attentiveness/ability to tolerate boredom”.

Importance of vertical scaling

“Educational measurement practitioners commonly violate the basic principle: “When measuring change, don’t change the measure.” Unlike instruments for measuring physical characteristics (e.g., height or weight), our instruments for measuring status and growth in educational achievement must change in order to preserve the validity of the measurements themselves.”

--From Patz (2007) *Vertical Scaling in Standards-Based Educational Assessment and Accountability Systems*. Published by Council of Chief State School Officers.

2. Decide on a metric for time

- In most studies we can define time in many ways.

1. Child welfare study.
   - Sequential recruitment. Thus time scaling options are:
     - Time from project start (Oct 2013).
     - Time from entry into DCF system.
       - Case is opened but screening, etc. occurs before enrollment into study.
     - Time from client enrollment/randomization (seems like best choice to me).

More examples

2. Studies of school children.
   - Chronological age.
   - School grade.
3. Studies of psychotherapy.
   - Weeks since entry into therapy.
   - Number of sessions.

Different outcome objectives=different time metrics

(Singer and Willett, 2003)

- IES RFA tells us to get measures sensitive to change.
- Here we want time metrics sensitive to change.

Imagine studying cars:

- Assess factors impacting appearance (rust, seat wear) → time since manufacture.
- Assess factors impacting "general wear and tear" (tire tread, belts) → miles driven.
- Assess the starter/ignition → trips driven.

3. Decide on number and spacing of measurements

? Equally spaced measurement occasions.

? Time structured (everyone measured at same values of time).

- ▪ More important for SEM or ANOVA analyses than the HLM analyses we will discuss here.

? Balanced (the same number of measurements taken on each person).

- You may not be able to totally control this.

- How you define time can impact whether data is considered time-structured or not.

Example:

- Child welfare study will "capture" data at 6 month intervals from beginning of study.
- If we count time from study onset → data would be time-structured.
- If we count time from client enrollment (better idea) → data would be time-unstructured.

3. Growth Trajectories

The problem of fitting growth trajectories is complicated

It requires choosing a form for the growth trajectories

You should have a basic form in mind going into the analysis based on: (i) a theoretical/conceptual model and (ii) your study objectives (what you want to learn from growth model).

Do you just want to know if there is a “general upward trend”(linear model may be sufficient)?

Is it also important to know the rate of growth at different time points (might need quadratic or cubic growth)?

General principle

Occam's razor

- All other things equal, Simpler is better.
- More complicated model= harder to estimate precisely=larger sample size required to estimate parameters precisely.
- Additional random effects (and covariance parameters) in model make this especially true.

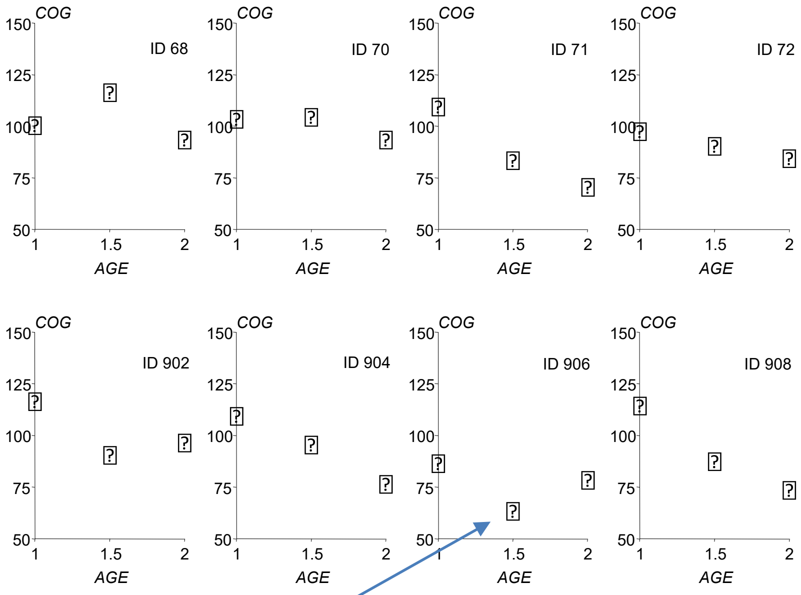
In other words, you might want to default to a linear or quadratic model and be parsimonious in adding random slope parameters.

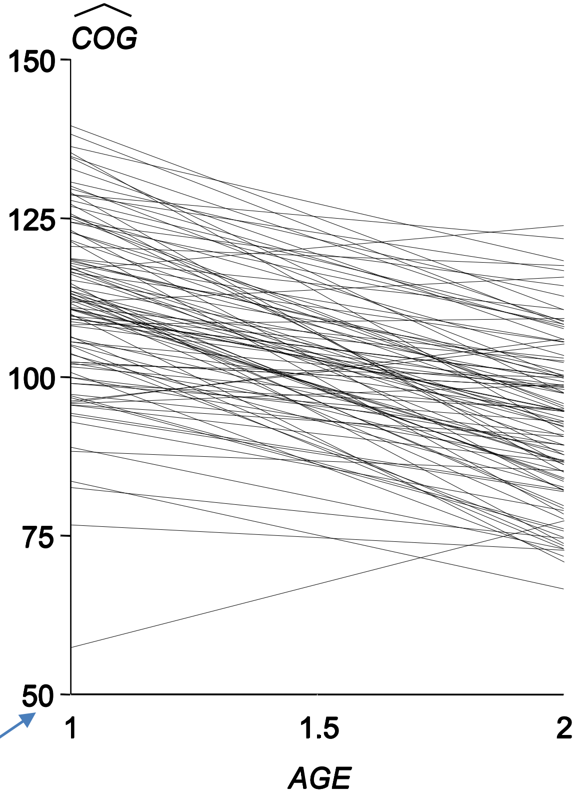
Possibly, be willing to modify form of the model based on an initial graphical exploration of the data.

- However, this is not as easy as it may seem.

Initial Graphical explorations: Make some plots like this

(adapted from Singer and Willet Applied Longitudinal Data Analysis)





Do we need a quadratic effect? Or is it just noise?

Do we need randomly varying slopes? Covariance between slope and initial status?

(ALDA, Section 3.2, pp. 49-51)

What to make of this?

- Very hard to say how much of what you see in these plots is “signal” and how much is “noise”.
- My advice: Only deviate from initial plan if evidence is fairly overwhelming that you need to.
  - For “pre-registration of protocol” reasons if for nothing else.

- Best not to focus on “exploratory aspects” of experimental study.
  - Eg. Extensive data exploration to determine the shape of a growth curve
  - These explorations should be done prior to experiment so you come into study with pretty good idea of model that you want to compare across the groups randomized.
- Experiments help with causal attribution.
- They <u>don’t</u> help us to build a good model for anything not having to do with treatment assignment.

Growth Trajectories (forms)

Many forms are possible, but polynomials are conventional for two reasons:

- Any smooth function is approximately a polynomial (Taylor's Theorem).
  - And approximately linear if you examine function over short enough period of time.
- Polynomials are simple.
  - In particular, they are linear in the parameters, which allows Inference based on normal theory.

What is a Polynomial Model?

$Y_{ijk} = \beta_{0jk} + \beta_{1jk}t_{ijk} + \beta_{2jk}t_{ijk}{}^{2} + \beta_{3jk}t_{ijk}{}^{3} + \ldots + \varepsilon_{ijk}$

t_{ijk} is defined as a measure of time for the jth person in the kth school on the ith measurement occasion.

Rarely do we go beyond a cubic function.

"jk" subscript implies each person has unique growth curve. We want to model these curves and the variation in these curves.

Note that the measurements do not have to be at exactly the same time for each person (time-unstructured OK).

How do we interpret these coefficients?

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} t_{ijk} + \beta_{2jk} t_{ijk}{}^{2} + \beta_{3jk} t_{ijk}{}^{3} + \ldots + \varepsilon_{ijk}$$

- β_{1jk} tells us the linear growth rate.
- β_{2jk} tells us the quadratic growth rate.
- β_{3jk} tells us the cubic growth rate.
- This is true, but not necessarily helpful. How do we interpret a linear growth rate when there are also quadratic and cubic terms?

Centering

- We typically center the measurements at some point for convenience
  - Common choices are:
  1. Middle
  2. Beginning
  3. End

- Centering strategy determines the interpretation of the coefficients of the growth model.

Understanding a Polynomial Model

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}t_{ijk} + \beta_{2jk}t_{ijk}^{2} + \beta_{3jk}t_{ijk}^{3} + \varepsilon_{ijk}$$

How do we interpret the coefficients?

β_{0jk} is the intercept at the centering point

β_{1jk} is the linear rate of growth at the centering point

- It is not the linear rate of growth anywhere else.

β_{2jk} is the acceleration (rate of change of linear growth) at the centering point.

- It is not the acceleration anywhere else.

β_{3jk} is the rate of change of the acceleration (often negative leading to a gradual flattening out of the growth curve).

Why cubics so important?

- Negative cubic term ensures growth doesn’t explode.
- Many natural growth processes have a “S” shape.
- We can produce this basic shape with a cubic.

Understanding a Polynomial Model

Consider the quadratic growth model to understand **changes in** growth **rate** with mean centering:

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + \beta_{1jk}(t - \bar{t}) + \beta_{2jk}(t - \bar{t})^2 \\ &= \beta_{0jk} + [\beta_{1jk} + \beta_{2jk}(t - \bar{t})](t - \bar{t}) \end{aligned}$$

Thus you can see that the linear growth rate at time t is

$$[\beta_{1jk} + \beta_{2jk}(t - \bar{t})]$$

In other words, the linear growth rate increases with t (assuming β_{2jk} is positive) and the only place where the linear growth rate is β_{1jk} is the middle.

NB: We've rewritten quadratic growth model as a linear growth model where rate of growth depends on *t*.

Understanding a Polynomial Model

Thus β_{1jk} is the linear rate of growth at the centered value (here, the middle)

If $\beta_{2jk} > 0,$ the linear growth rate will be larger above the centered value and smaller below the centered value

Centering at other values than the middle can make sense if that is where growth trajectory is of interest and if the model fits the data.

- Eg. Coding “0”= first time point, “1”=second time point, … is centering at the beginning.

For example, centering at the end gives coefficients with interpretable rates at the end of the growth period.

- Eg. You care most about how kids are doing at the end of the study.

Understanding a Polynomial Model

Consider the cubic growth model to understand **changes in the acceleration of the growth rate** with mean centering

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + \beta_{1jk}(t-\bar{t}) + \beta_{2jk}(t-\bar{t})^2 + \beta_{3jk}(t-\bar{t})^3 \\ &= \beta_{0jk} + \{\beta_{1jk} + [\beta_{2jk} + \beta_{3jk}(t-\bar{t})](t-\bar{t})\}(t-\bar{t}) \end{aligned}$$

Thus you can see that the *acceleration* at time t is

$$[\beta_{2jk} + \beta_{3jk}(t-\bar{t})]$$

In other words, the acceleration increases (decreases if β_{3jk} is negative) with t and the only place where the acceleration is β_{2jk} is the middle

Understanding a Polynomial Model

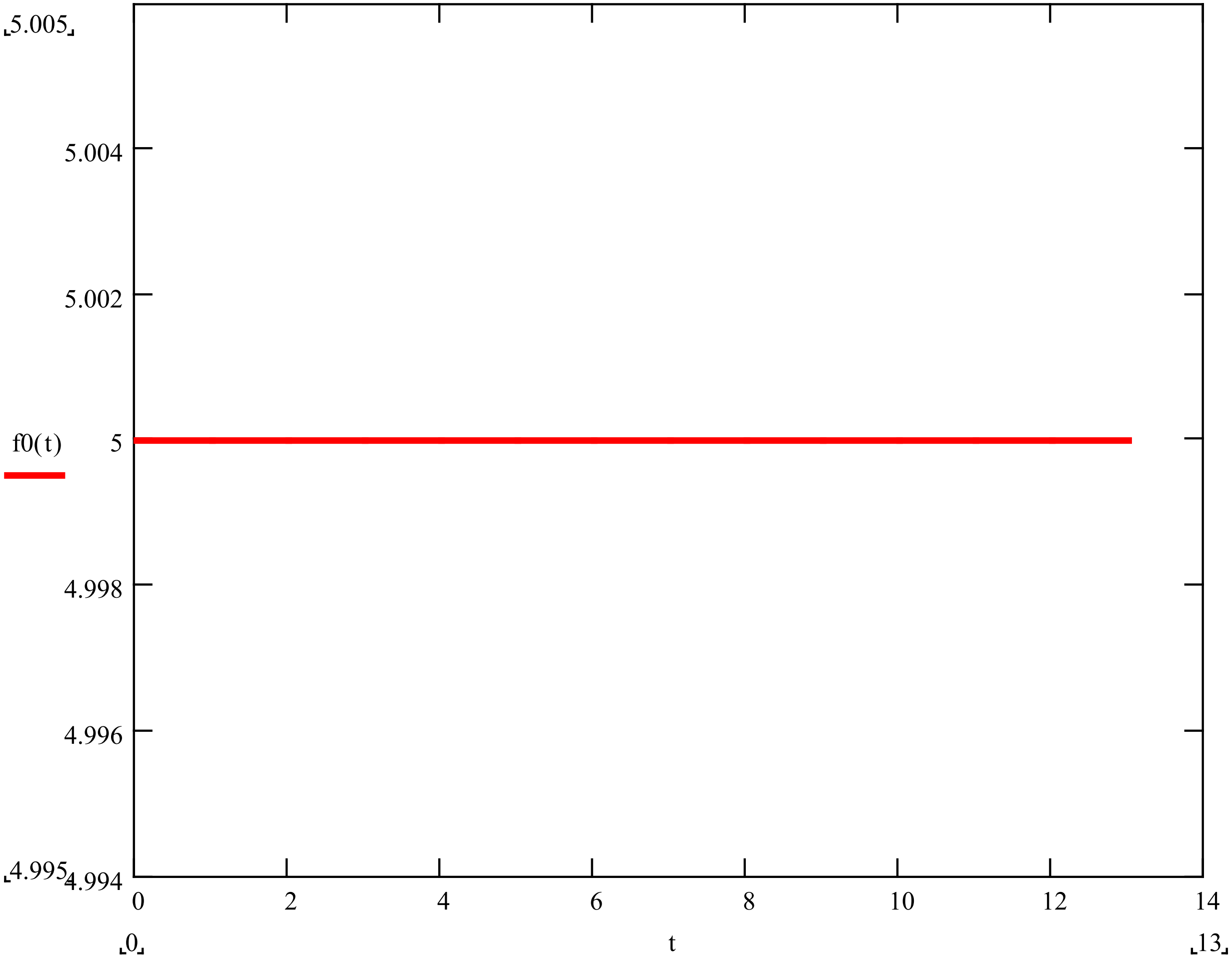
Thus β_{2jk} is the acceleration of growth at the centered value (near the middle)

If $\beta_{3jk} < 0,$ the acceleration will be larger below the centered value and smaller above the centered value

Again, centering at other values than the middle can make sense if that is where growth trajectory is of interest and if the model fits the data

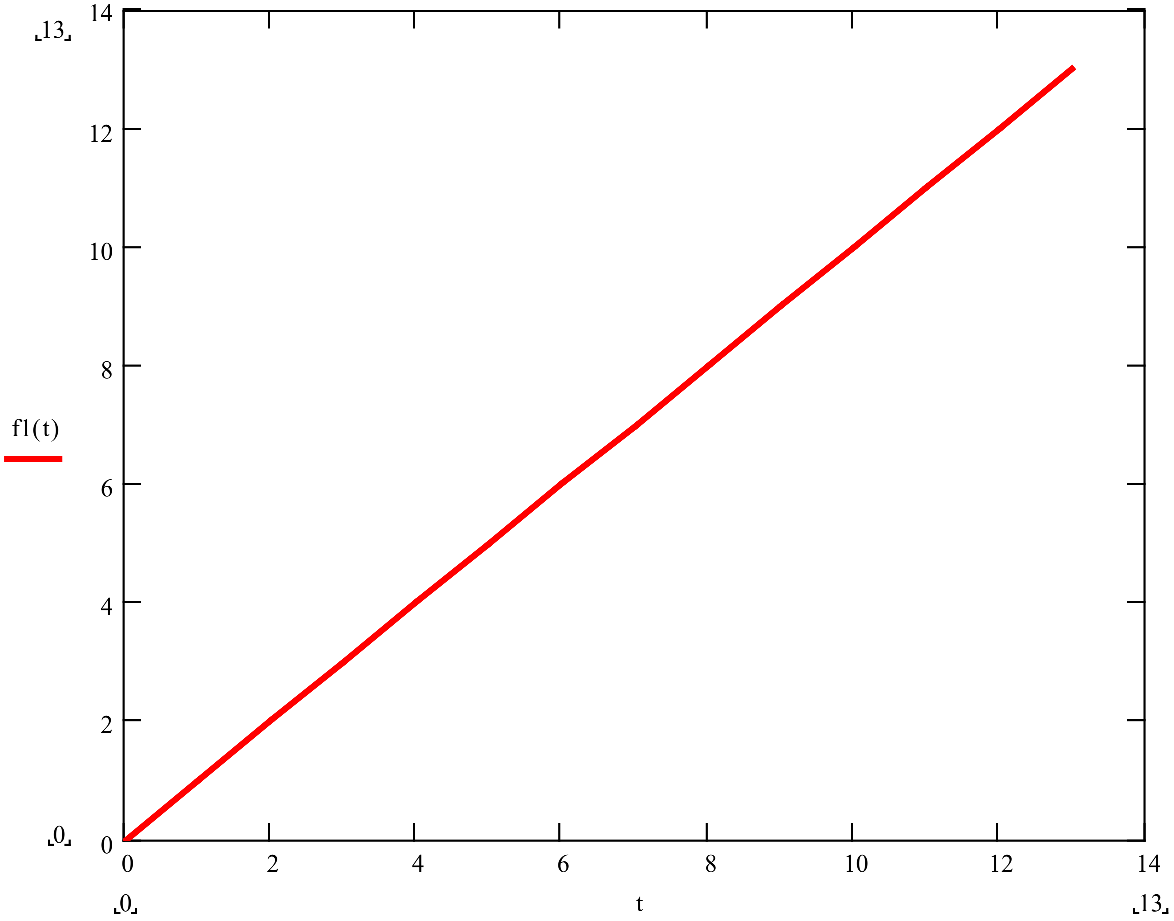
No Growth (centering irrelevant)

$$\beta_0 = 5, \beta_1 = 0.00, \beta_2 = 0.00, \beta_3 = 0.00$$



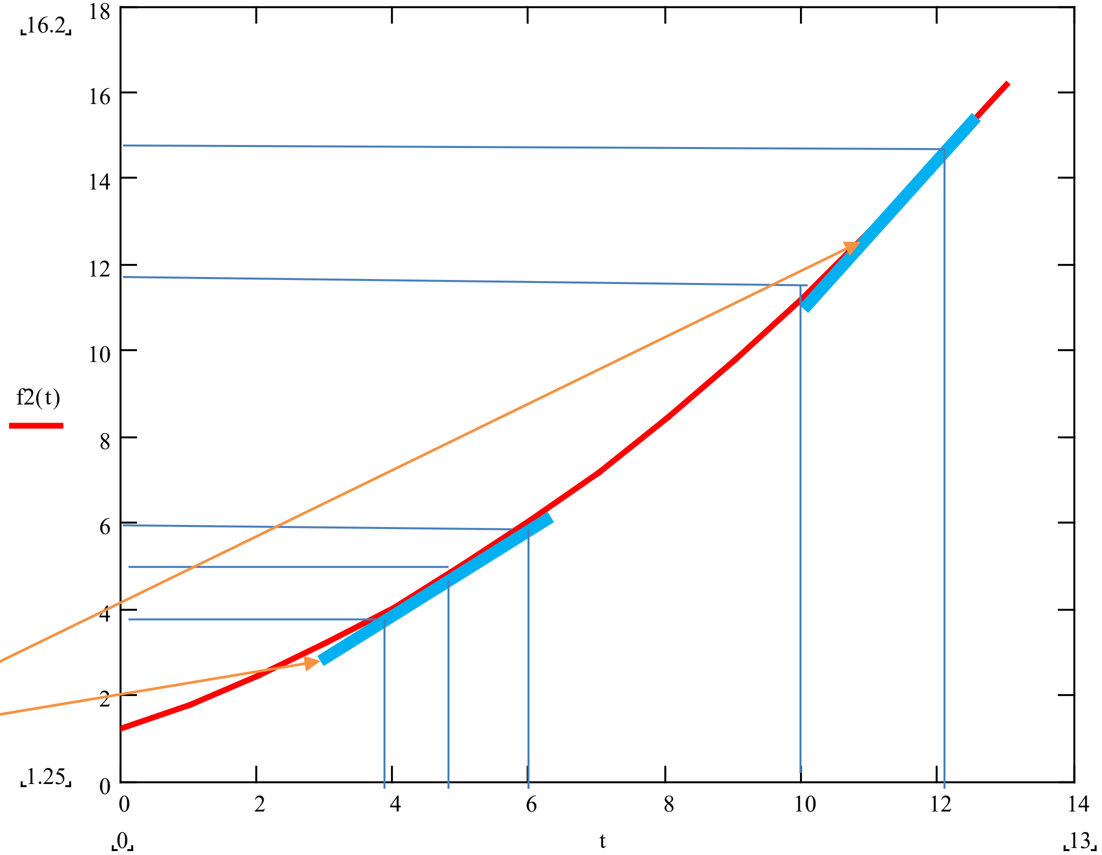
Linear Growth (centered at 5)

$$\beta_0 = 5, \beta_1 = 1, \beta_2 = 0.00, \beta_3 = 0.00$$



Quadratic Growth (Centered at 5)

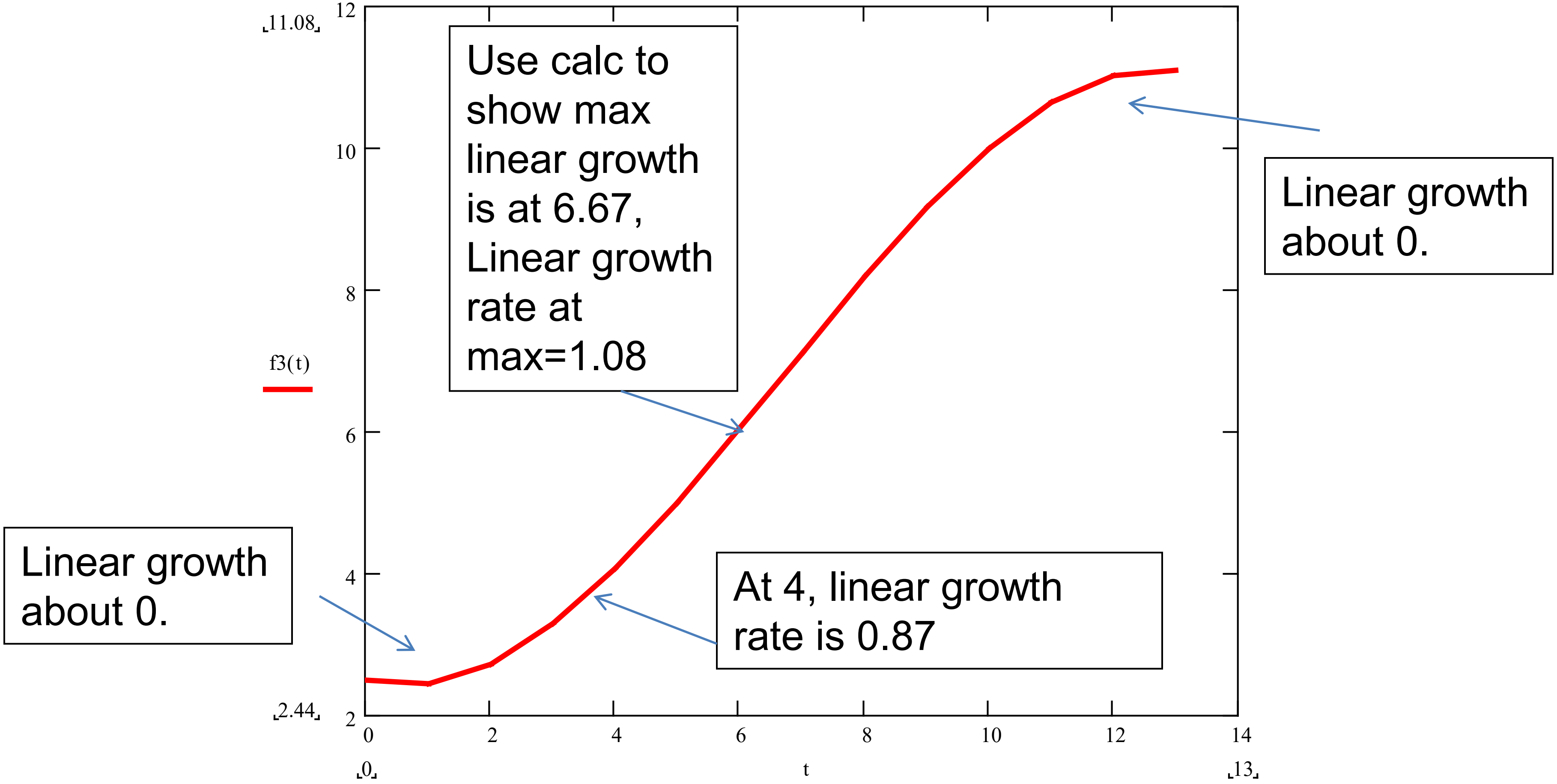
$$\beta_0 = 5, \beta_1 = 1, \beta_2 = 0.05, \beta_3 = 0.00$$



Tangent lines

Cubic Growth (Centered at 5)

$$\beta_0 = 5, \beta_1 = 1, \beta_2 = 0.05, \beta_3 = \text{-}0.01$$



Use calc to show max linear growth is at 6.67, Linear growth rate at max=1.08

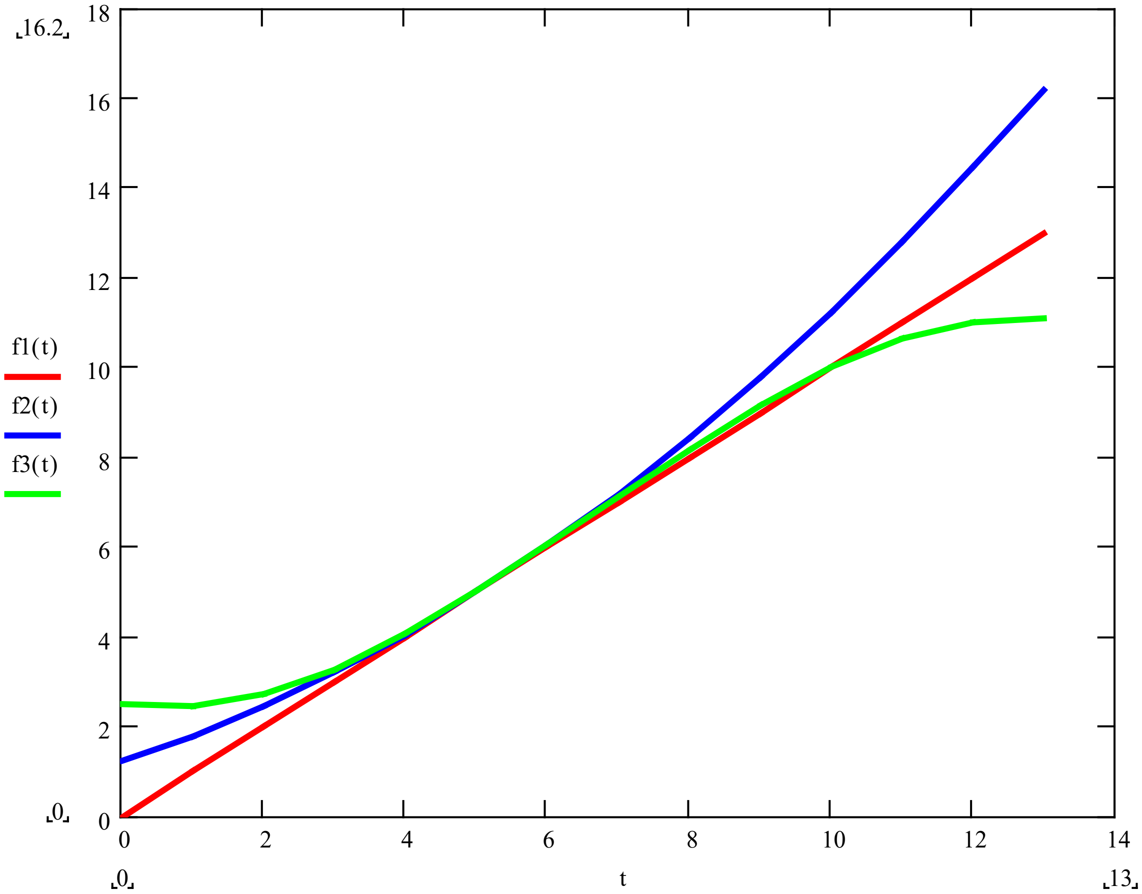
Linear growth about 0.

Linear growth about 0.

At 4, linear growth rate is 0.87

Linear, Quadratic, and Cubic Growth

$\beta_0 = 5, \beta_1 = 1, \beta_2 = 0.05, \beta_3 = -0.01,$



Notice: Near the centering point all growth is linear

*So you have maximal robustness to mis-specified model at centering point.
* You also have maximum precision of estimated of fitted value

Selecting Growth Models

Several considerations are relevant in selecting a growth model

First is how many repeated measures there are: The maximum degree is one less than the number of measures

(linear needs 2, quadratic needs 3, etc.)

However the estimates of growth parameters are much better if there are a few additional degrees of freedom

But the most important considerations are:

(i) Research questions.

(ii) Whether the model fits the data!

- Unfortunately, as explained earlier, making this determination is not that easy.

Selecting Growth Models

Individual growth trajectories are usually poorly estimated

- Even with as many as 8 time points, to estimate cubic growth is like estimating a multiple regression with 4 parameters from 8 data points for each person.
- More data points would help, but they can be expensive.

HLM models estimate average growth trajectories (via average parameters) and variation around that average: These are much more stable.

Selecting Growth Models

- Usually bad idea to let empirical data entirely guide the choice of the model.
  - There is too much noise to let data guide you in its entirety. There will usually be multiple curves that fit just about equally well.
  - Let both theory, research questions and data guide your model choice.

Advice:

1. If your initial (normative, theory based) choice seems an extremely poor fit, revise model.
2. However, one need not be a slave to “fit statistics” that may be sensitive to distributional assumptions that may not be met.
3. Overfitting (modeling noise) and underfitting (missing an essential structural feature of the curve) are both dangers.

Estimating and Interpreting individual growth curves

Estimates of individual growth curves can usually be greatly improved by using empirical Bayes methods to borrow strength from the averages.

* “shrinkage” estimates which are effectively weighted average between individual growth curve and average growth curve.

This makes the most sense if all the individuals in the groups are sampled from a common population.

It can be problematic if some individuals are dramatically different.

– But, hard to know if this is the case just by looking at data, for reason just mentioned (noise in the data).

Implications of heterogeneity in individual growth curves

- IF there is a lot of heterogeneity how do we interpret the “average” curves that we estimate?

Implications of heterogeneity in individual growth curves

- No easy answer, however:

1. These are the situations where it may be quite interesting to look at the covariance between the random effect terms.
   - E.g. do the kids that have high linear growth (high growth near the centering point) also have lower quadratic terms?
     - Is there a “catching up” effect? OR
     - If you fall way behind are you doomed?

Back to Experiments

- However, things like:

1. Model selection (randomization can't help us choose the right model).
2. Interpreting covariance parameters

are ideally <u>not</u> the main foci of experiments.

EXCEPT

- We "might" be interested in how treatment influences the covariance parameter.

Selecting Analysis Models

One issue is selecting the growth model to characterize growth

A different, but related, issue is selecting how treatment should impact growth

Should it impact linear growth term?

Should it impact the acceleration?

Which impact is primary? (e.g. for power and secondary/primary research question purposes).

How does looking at multiple impacts weaken the design?

- multiple comparisons.

What if impacts are in opposite directions?

- How do we interpret and report this?

Please show me a slide with lots of Greek letters that shows what all this growth modeling looks like in HLM, you ask?

I would not want to disappoint…

Longitudinal Experiments Assigning Treatment To Schools (quadratic model)

Level 1 (measures)

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} t_{ijk} + \beta_{2jk} t_{ijk}^{2} + \varepsilon_{ijk}$$

Level 2 (individuals)

$$\begin{aligned} \beta_{0jk} &= \gamma_{00k} + \eta_{0jk} \\ \beta_{1jk} &= \gamma_{10k} + \eta_{1jk} \\ \beta_{2jk} &= \gamma_{20k} + \eta_{2jk} \end{aligned} \qquad \boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{I})$$

Level 3 (schools)

$$\begin{aligned} \gamma_{00k} &= \pi_{000} + \pi_{001} T_k + \xi_{00k} \\ \gamma_{10k} &= \pi_{100} + \pi_{101} T_k + \xi_{10k} \\ \gamma_{20k} &= \pi_{200} + \pi_{201} T_k + \xi_{20k} \end{aligned} \qquad \boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{S})$$

NB : If time centered at the beginning (time of randomization), π_{001} should be 0 in experiments. Why?

Longitudinal Experiments Assigning Treatment To Schools (quadratic)

This model has three trend coefficients in each growth trajectory

Note that there are 3 random effects at the second and third level

This means that 6 variances and covariances must be estimated at each level

This may require more information to do accurately than is available at the school level

It is often prudent to fix some of these effects because they cannot all be estimated accurately.

Longitudinal Experiments Assigning Treatment Within Schools (quadratic)

Level 1 (measures level)

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk}\, t + \beta_{2jk}\, t^2 + \varepsilon_{ijk} \qquad \varepsilon \sim \mathrm{N}(0,\ \sigma_W{}^2)$$

Level 2 (individual level)

$$\beta_{0jk} = \gamma_{00k} + \gamma_{01k}T_j + \eta_{0jk} \qquad \boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0},\ \boldsymbol{\Sigma}_C)$$
$$\beta_{1jk} = \gamma_{10k} + \gamma_{11k}T_j + \eta_{1jk}$$
$$\beta_{2jk} = \gamma_{20k} + \gamma_{21k}T_j + \eta_{2jk}$$

Covariance structure for control group schools (if "T_j" is "0" or "1" coded).

Level 3 (school level)

$$\gamma_{00k} = \pi_{000} + \xi_{00k}$$
$$\gamma_{01k} = \pi_{010} + \xi_{01k}$$
$$\gamma_{10k} = \pi_{100} + \xi_{10k}$$
$$\gamma_{11k} = \pi_{110} + \xi_{11k}$$
$$\gamma_{20k} = \pi_{200} + \xi_{20k}$$
$$\gamma_{21k} = \pi_{210} + \xi_{21k}$$

$$\xi_{a0} \sim \mathrm{N}(\mathbf{0},\ \boldsymbol{\Sigma}_S)$$
$$\xi_{a1} \sim \mathrm{N}(\mathbf{0},\ \boldsymbol{\Sigma}_{TxS})$$

Covariance structure of treatment effects.

Longitudinal Experiments Assigning Treatment Within Schools

This model has three trend coefficients in each growth trajectory

Note that there are 6 random effects at the third level

This means that 15 variances and covariances must be estimated at the third level

This requires a great deal of information to do accurately

It is often prudent to fix some of these effects because they cannot all be estimated accurately

However there is some art in this, and sensitivity analysis is a good precaution

Covariates

Covariates can be added at any level of the design

But remember that covariates must be variables that cannot have been impacted by treatment assignment

Thus time varying covariates (at level 1) are particularly suspect since they may be measured *after* treatment assignment.

- By "suspect" I mean their inclusion will preclude interpreting model coefficients as unbiased causal effects.

Time varying “covariate” example

- Intervention involves different forms of psychotherapy.
- Depressive symptoms measured after each session.
- Case #1: some sessions scheduled for 45 minutes, others for 1 hour.
- Case #2: some sessions “Cut short” due to patient or doctor request (so some are 45 min, others 1 hour).
- How does distinction impact our use of the variable?

And Another Example

- Students changing schools during the course of the study.
- Should we control for school they are attending at post-test (as opposed to school attending when randomized)?
  - E.g. add fixed or random effect of this school to our models?

Summary of growth modeling good practices

1. Make sure measure used appropriate for studying growth.
   - Vertical scaling.
2. Should have some theoretical/conceptual model that helps pick functional form of growth model.
3. Be willing to modify initial model based on visual inspection of data.
   - Although frequently this will be uninformative.

Summary of growth modeling good practices

4. Compare model predicted values to actual values.
   - Make sure not too discrepant.
5. Make sure you have enough data to estimate the model you specify.
   - If not, fit a simpler model.
   - Ideally, realize this ahead of time and don’t waste resources on the wrong things.
     - ie. more measures vs. more schools or subjects.

- Experiments are expensive.
- Worth it (we hope) because of ability to produce clear causal conclusions.
  - We can give observed differences b/tw T and C causal interpretations.
- Ideally we would NOT cloud our ability to observe treatment effects.

a) By trying to estimate models that are too complex to estimate well.

b) By analyzing data contaminated by attrition.

If time

WWC and growth modeling

WWC and growth modeling

- What works clearinghouse (WWC) has standards for evaluating educational studies.
  - Low attrition RCTs and certain RD studies are only studies that can meet standards without reservations.
- What does WWC say about growth models?

WWC and growth models

- *“growth curve analyses do not typically provide point-in-time impact estimates. However, the WWC will request the data needed from authors to calculate effect sizes—and baseline equivalence, if required—at each point in time (WWC standards version 4.1, p. 32).”*
  - In other words, WWC is going to take your growth curve analysis and do its best to reconstruct what estimate would have been at each time point separately.

If time: Revisit running example

What sorts of things can go wrong as educational RCTs move into Y2 and Y3 of implementation?

Scenario #1:

Three 3rd grade cohorts (same 3rd grade teachers). D and E index cohorts. Could add 4th (measure level) to account for three measurement per students.

Level 1 (student level)

$$Y_{ijk} = \beta_{0jk} + \beta_{Djk}D_{ijk} + \beta_{Ejk}E_{ijk} + \varepsilon_{ijk} \qquad \varepsilon \sim \mathrm{N}(0, \sigma_W^2)$$

Level 2 (teacher level)

$$\beta_{0jk} = \gamma_{00k} + \eta_{0jk}$$

$$\boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma_I})$$

$$\beta_{Djk} = \gamma_{D0k} + \eta_{Djk}$$

$$\beta_{Ejk} = \gamma_{E0k} + \eta_{Ejk}$$

Level 3 (school level)

$$\gamma_{00k} = \pi_{000} + \pi_{001}T_k + \xi_{00k}$$

$$\boldsymbol{\xi} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma_S})$$

$$\gamma_{D0k} = \pi_{D00} + \pi_{D01}T_k + \xi_{D0k}$$

$$\gamma_{E0k} = \pi_{E00} + \pi_{E01}T_k + \xi_{E0k}$$

Scenario #1: But…

- Suppose:

1. Not all teachers in every school consent to be in study.
2. In years 2 and 3 of study principals in T schools decide to put their struggling students in the classrooms of participating teachers.

Scenario #2

- Follow students into Y2 and Y3.
- But, since not all teachers consented, some students get teachers exposed to PD in G3, some in G3+G4, some in G3+G5, some in G3+G4+G5.
- What to do?
- How to interpret parameters?

If time

Power in longitudinal studies

Power Analysis

Power computations for longitudinal experiments are doable, but depend on parameters that may not be well known

For example reliability of trend coefficients.

- When instruments are psychometrically tested, almost always psychometrics are computed at a single time point.
- Vertical scaling is not, by itself, enough.
  - Just b/c we have vertical scale doesn’t always mean we have estimate of reliability of (e.g.) linear rate of change estimates.

Power Analysis

When parameters such as these are known, the computations are straightforward, but there is relatively little information about them that can be used for planning

To make matters worse, the values of some parameters (such as reliability) depend on the number of measures.

- So, just because you can do the power analysis for the case of three time points, doesn't mean you can (easily) do the analysis for the case of 4.

Power Analysis

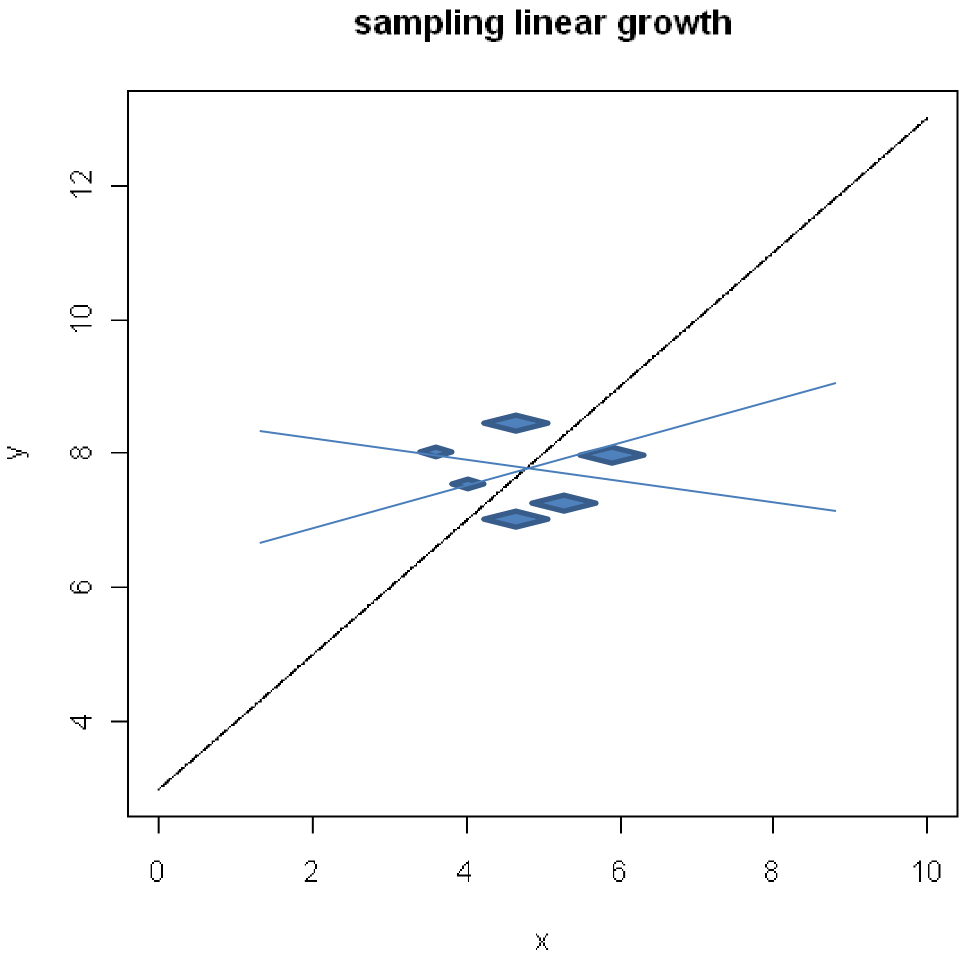
Still some generalizations are possible

- Power increases with the number of measures
- Power increases with the length of time over which measures are made (except for power for β_{0jk})
- Power increases with the precision of each individual measure

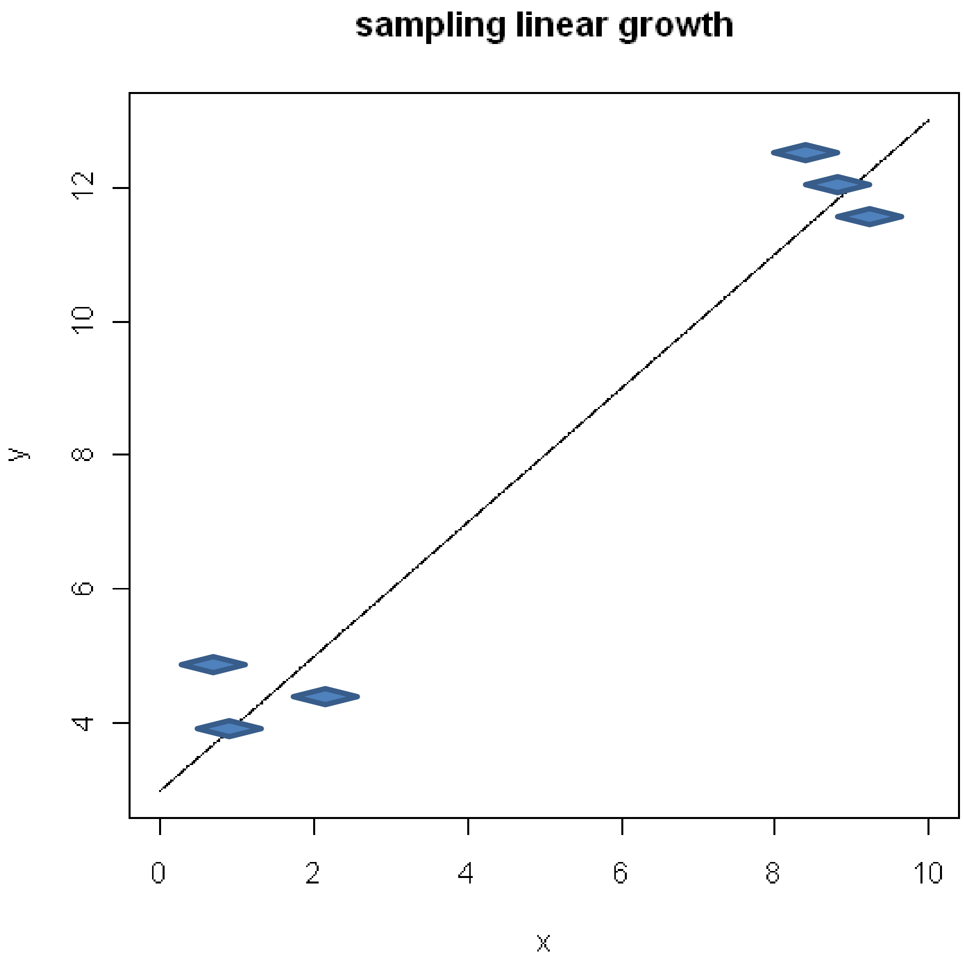
These factors impact different trend coefficients differently

Clustering increases the complexity of computations

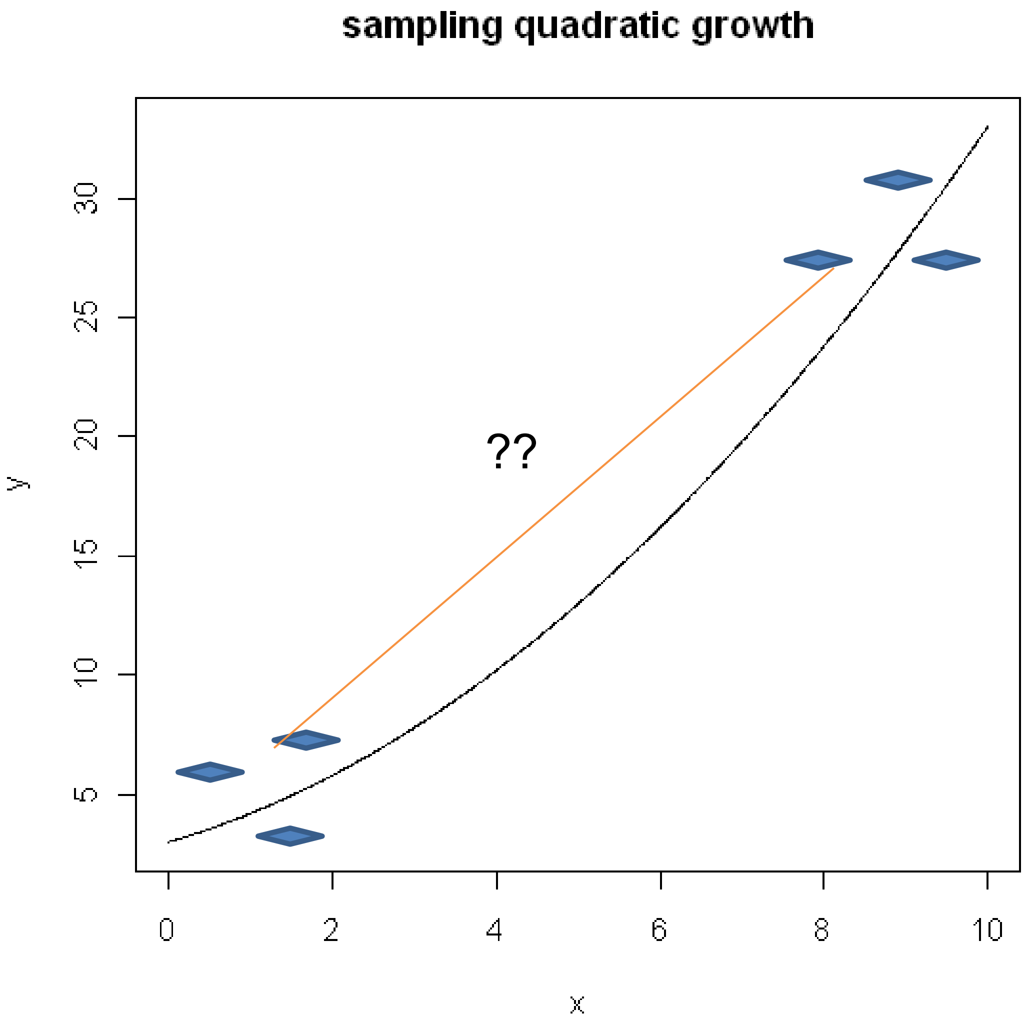
Power increases with the length of time



Power increases with the length of time



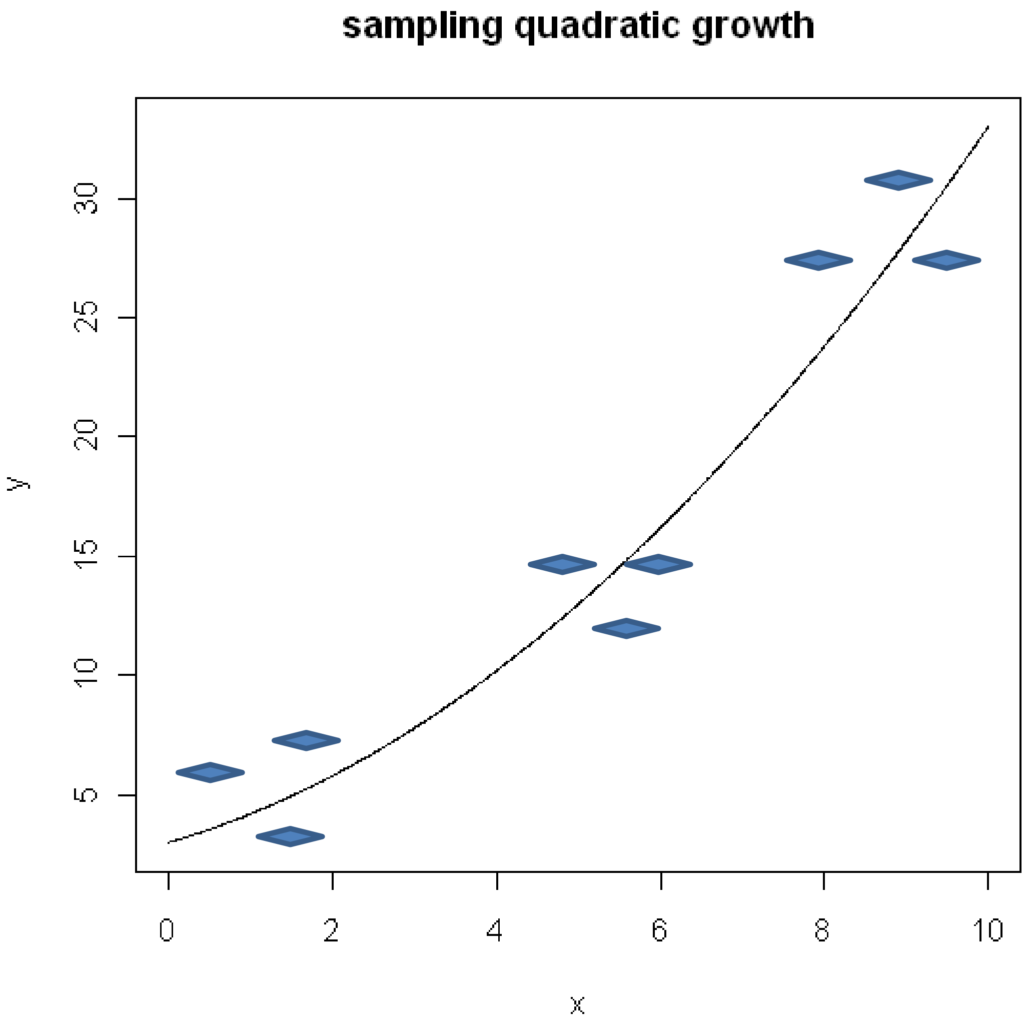
Impacts different coefficients differently



What is wrong here?

No robustness to incorrect model specification/ should quadratic term be 0?

Impacts different coefficients differently



Can’t get good curvature estimate on previous slides

Power Analysis

Pilot data (or data from related studies, perhaps non-experimental ones) is more important in planning longitudinal experiments than cross-sectional ones.

Because it is so important to get functional form right when computing power.

AND

Because it is hard to get good information about expected variation in things like slope coefficients.

Power

- Optimal Design can compute for person and cluster randomized trials (not directly for blocked trials).

Relevant parameters:

- D=Duration
- *f*=frequency of observation
- Variation of level 1 coefficient:
- Variation of level 1 residual:
- Effect size

These are parameters that can be very hard to determine.

Power example(s)

- Li and Konstantopoulos (2019) extend existing work to the case of block randomized trials.
  - Specifically, they consider a three level design (e.g., measurements, students, schools) where students within schools are assigned to conditions.
  - They assume equally spaced time points and one measurement per unit time.
    - i.e. don’t distinguish between frequency and duration, unlike OD/Raudenbush and Liu (2001).

Power examples
Li and Konstantopoulos (2019)

- For any trend coefficient (e.g. linear, quadratic) power depends on:
  - # measurement occasions
  - # students per school
  - # schools
  - level 1 measurement error.
  - Level 2 variation in trend coefficients.
  - Level 3 variation in treatment impacts on trend coefficients.
  - Effect size (avg. difference b/tw T and C for that coefficient) standardized in some fashion.

Power examples Li and Konstantopoulos (2019)

- They standardize effect size by square root of sum of level two variance in trend coefficient plus level three treatment effect variance.
  - Wouldn’t make sense to include measurement error in denominator.
- Using parameters from Project STAR, they find that with 4 measurement occasions, 40 schools and 30 students per school:

1. Power for linear effect size of .40 is 0.67
2. Power for quadratic effect size of .40 is 0.59.

More Extras if time

Missing data and binary outcomes

Missing data

- Some level of attrition is inevitable.

For estimating regression coefficients

- This is no problem.
- Modern software (like HLM) can easily estimate models with unbalanced, time-unstructured data.

Missing data

From standpoint of causal inference

- Potentially a big problem.

- We randomized to ensure equivalence (on average) between two groups on all factors besides treatment.
- Once there is attrition this guarantee is gone.

Differential attrition

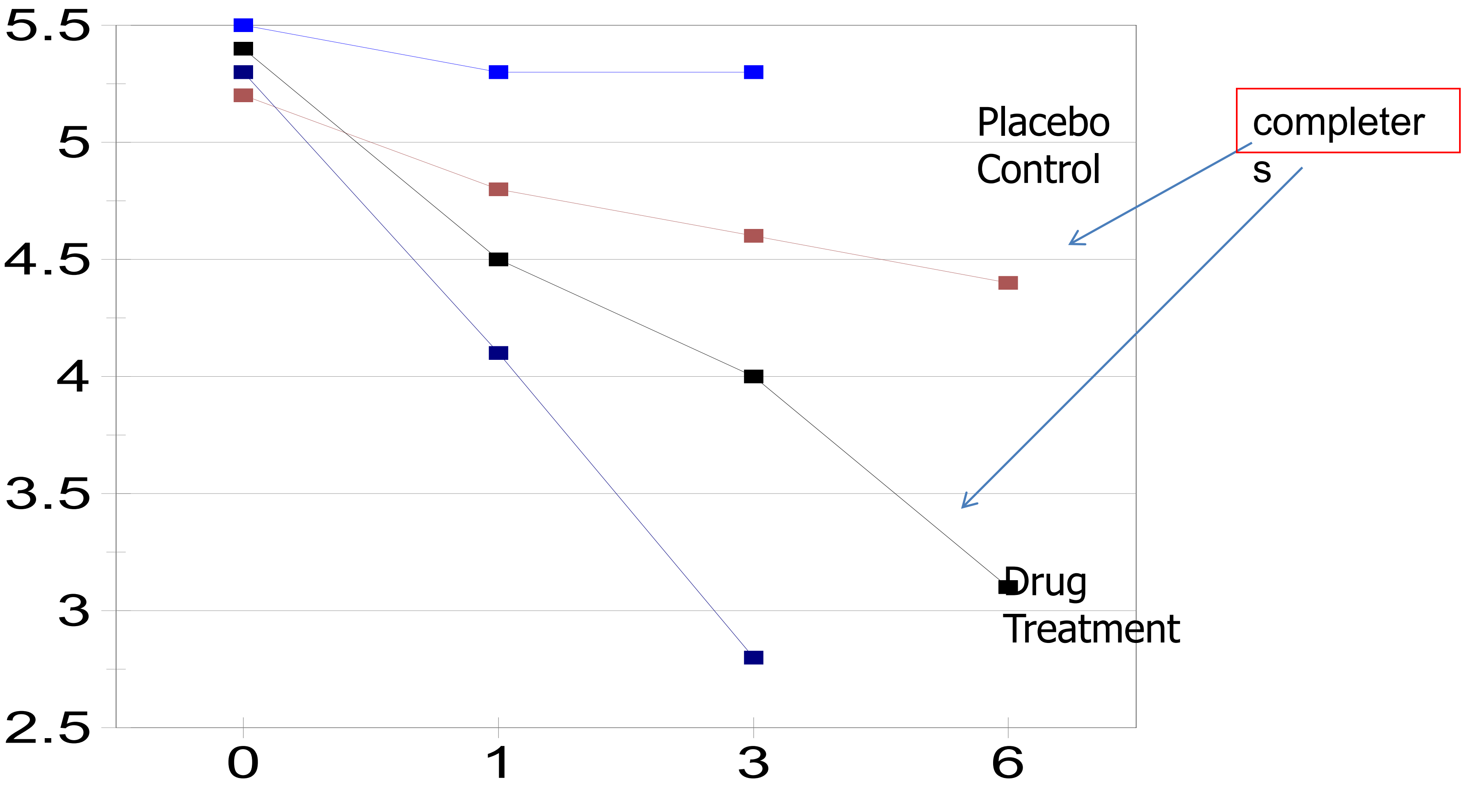
1. Less clear how to define differential attrition in longitudinal studies.
   - Perhaps 10% dropped out in T and C groups but T subjects mainly dropped after Time 2 and C subjects mainly after Time 4.
2. Lack of differential attrition does not guarantee lack of bias.
   - Just because the % that left study is the same does not mean those who left were “the same” or left for the same reasons.

Example: Hedeker and Gibbons (1997)

- Study of psychotherapy for depression.
  - In the treatment group those who left study are those who had been getting better fastest.
  - In the control group, those who left study are those who had been getting better slowest.
    - Graph in a moment.
  - Study could be salvaged, because they could document that the reasons for leaving depended on a measured variable.
  - However, if something similar occurred on an unmeasured variable causal inference would be seriously compromised.

From Hedeker & Gibbons (1997)

low = better outcomes



Weeks of Treatment

Binary outcomes

- Strictly speaking, the models presented today are appropriate only for <u>continuous</u> outcomes.
- However, analogous modeling approaches for dichotomous (binary) outcomes are available.
- Generally go by name “Generalized Linear Mixed Models” (GLMM).
  - The “logistic” and “probit” models are most common for binary data.

Binary outcomes

- Relationship to hierarchical linear models is analogous to relationship of linear regression to logistic regression.

- Conceptually, models can be formed in the same fashion as linear models.
  - That is, by adding “random” effects at higher levels to represent variation in logistic regression coefficients across higher level units.

Binary outcomes

- However, it is far more difficult to interpret the meaning of the parameters.
- In linear models we can think of decomposing variance into level-1, level-2, level-3 variance, etc.
  - This decomposition defines quantities like the ICC.
- An analogous definition of an ICC is not available in generalized linear models.

Binary outcomes

? Why not ?

- For binary variables the mean is related to the variance.
- Once you specify the mean structure of the model, you also specify the variance.
- So, the level 1 variance is fixed by assumption and cannot be estimated from the data.