

# AS TIME GOES BY...

## *An Introduction to the Analysis of Longitudinal Data*

Marie Davidian



davidian@stat.ncsu.edu

**Where to get a copy of these slides (and more):**

<http://www.stat.ncsu.edu/~davidian>

## Outline

1. Some examples and questions of interest
2. Some *ad hoc* approaches (and why they might not be so good...)
3. How do longitudinal data happen? – A conceptualization
4. Statistical models: Subject-specific and population-averaged
5. Methods for implementation
6. Discussion

## 1. Some examples and questions of interest

**Longitudinal studies:** Studies where a response is observed on each participant/unit *repeatedly over time* are commonplace, e.g.,

- Clinical trials, observational studies in humans, animals
- Studies of growth and decay in agriculture, chemistry

### Key messages in this talk:

- The *questions of interest* may be *different*, depending on the setting
- Longitudinal data have *special features* that must be taken into account to make *valid inferences* on questions of interest
- Statistical *models* and *methods* that acknowledge these features and the questions of interest are needed

## First, an “ideal” situation...

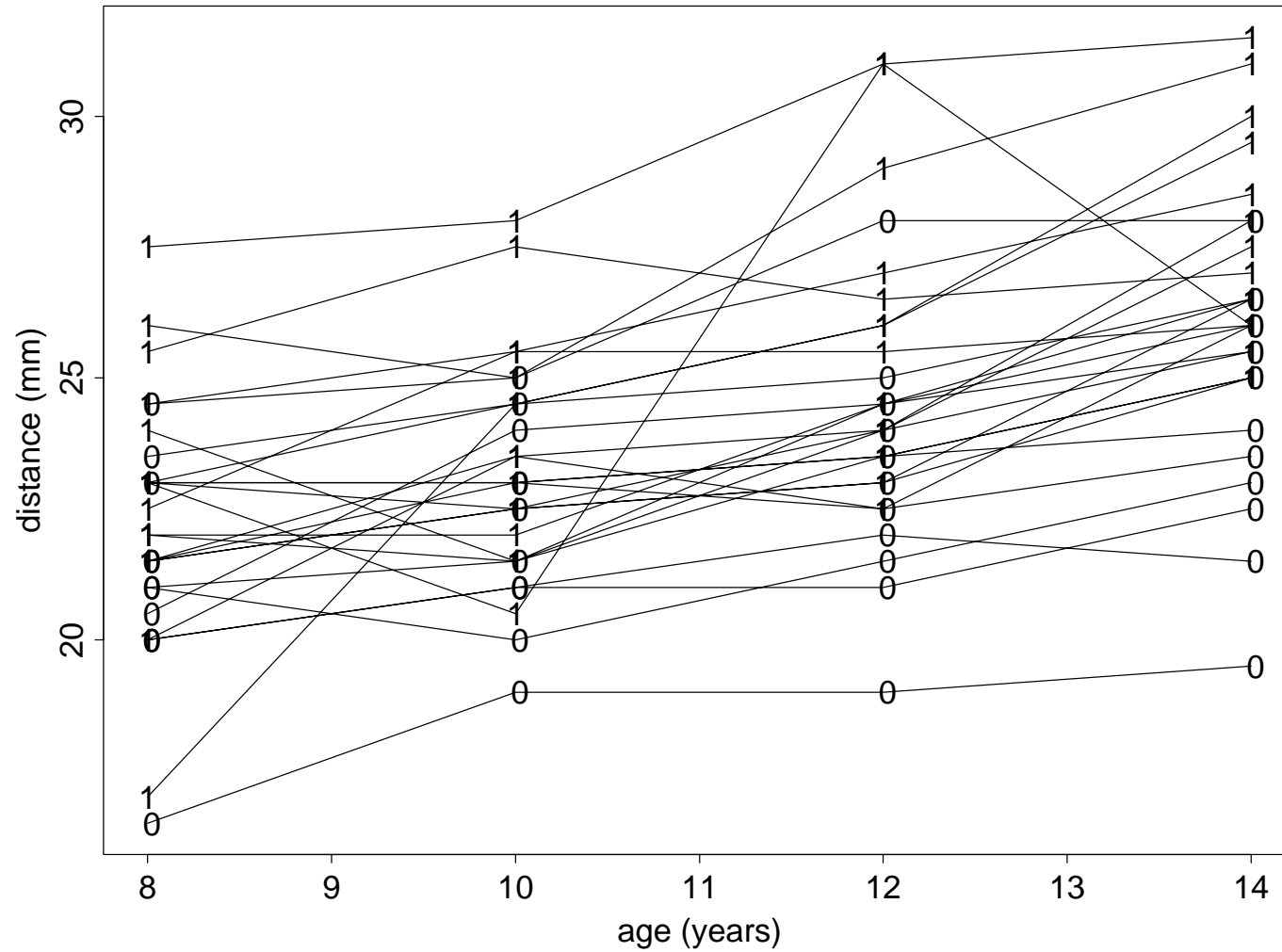
### “World-famous” dental study: Pothoff and Roy (1964)

- 27 children, 16 boys, 11 girls
- On each child, *distance* (mm) from the center of the pituitary to the pteryomaxillary fissure measured *on each child* at ages 8, 10, 12, and 14 years of age
- A *continuous* measure of growth

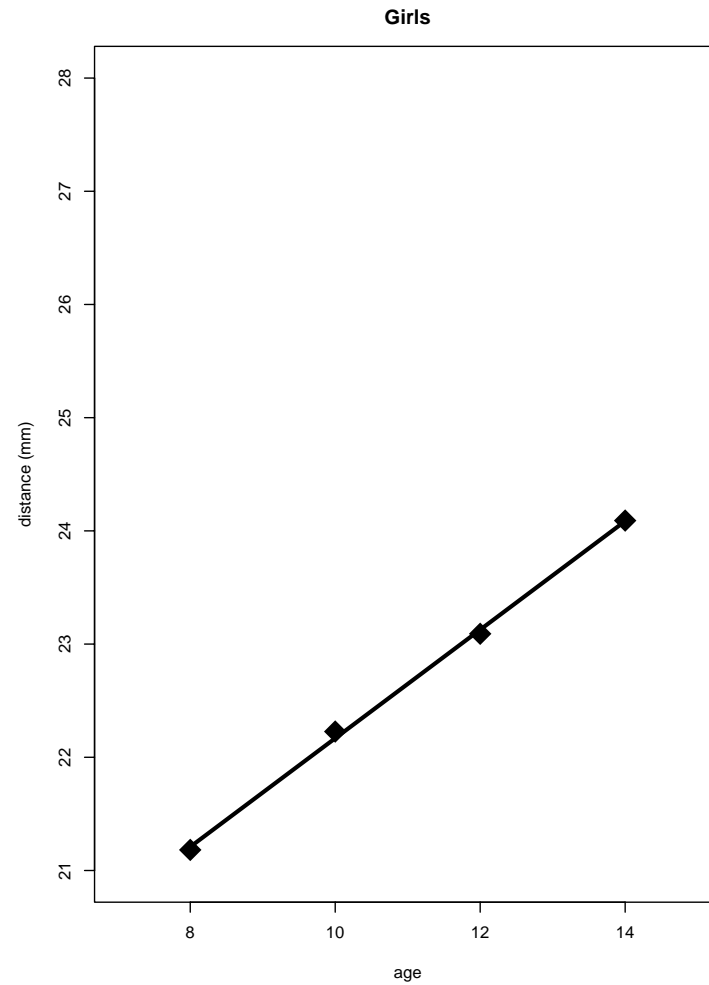
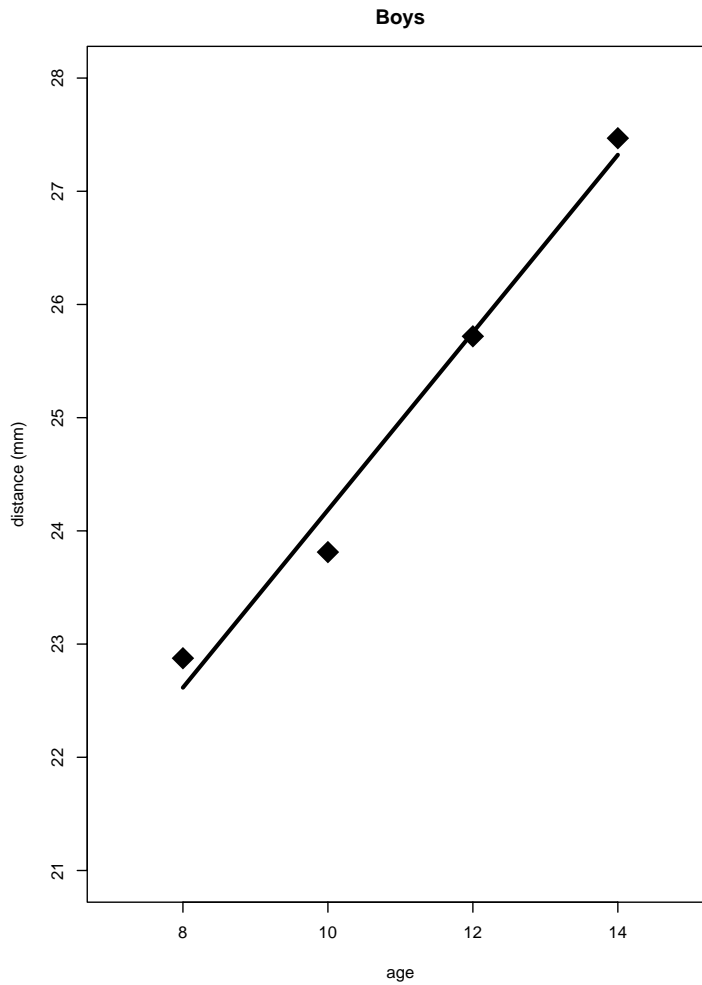
### Questions of interest: *Informally* stated

- Does distance *change* over time?
- What is the *pattern of change*?
- Is the pattern *different* for boys and girls? *How*?

All data (“spaghetti plot”): 0 = girl, 1 = boy



**Sample average dental distances:** Average across all boys, girls at each age



## Observations:

- *All children* have all 4 measurements at the same time points (ages) (“*balanced*”)
- Children who “*start high*” or “*low*” tend to “*stay high*” or “*low*”
- The *individual* pattern for *most children* follows a *rough straight line* increase (with some “*jitter*”)
- And *average distance* (across boys and across girls) follows an approximate *straight line* pattern

## Response need not be a continuous measurement...

### Another “famous” data set: Six Cities Study

- 300 children from six different cities examined annually at ages 9–12
- On each child, *respiratory status* (1=infection, 0=no infection) and *maternal smoking* in past year (1=yes, 0=no)
- *Discrete* (*binary*) response

### Questions of interest: *Informally* stated

- Is there an *association* between child’s respiratory status and mother’s smoking behavior?
- Does the association *change with age*?

### Observations:

- Graphical depiction not really informative (*binary* response)
- Further complication: *Missing* data at some ages for some mother-child pairs

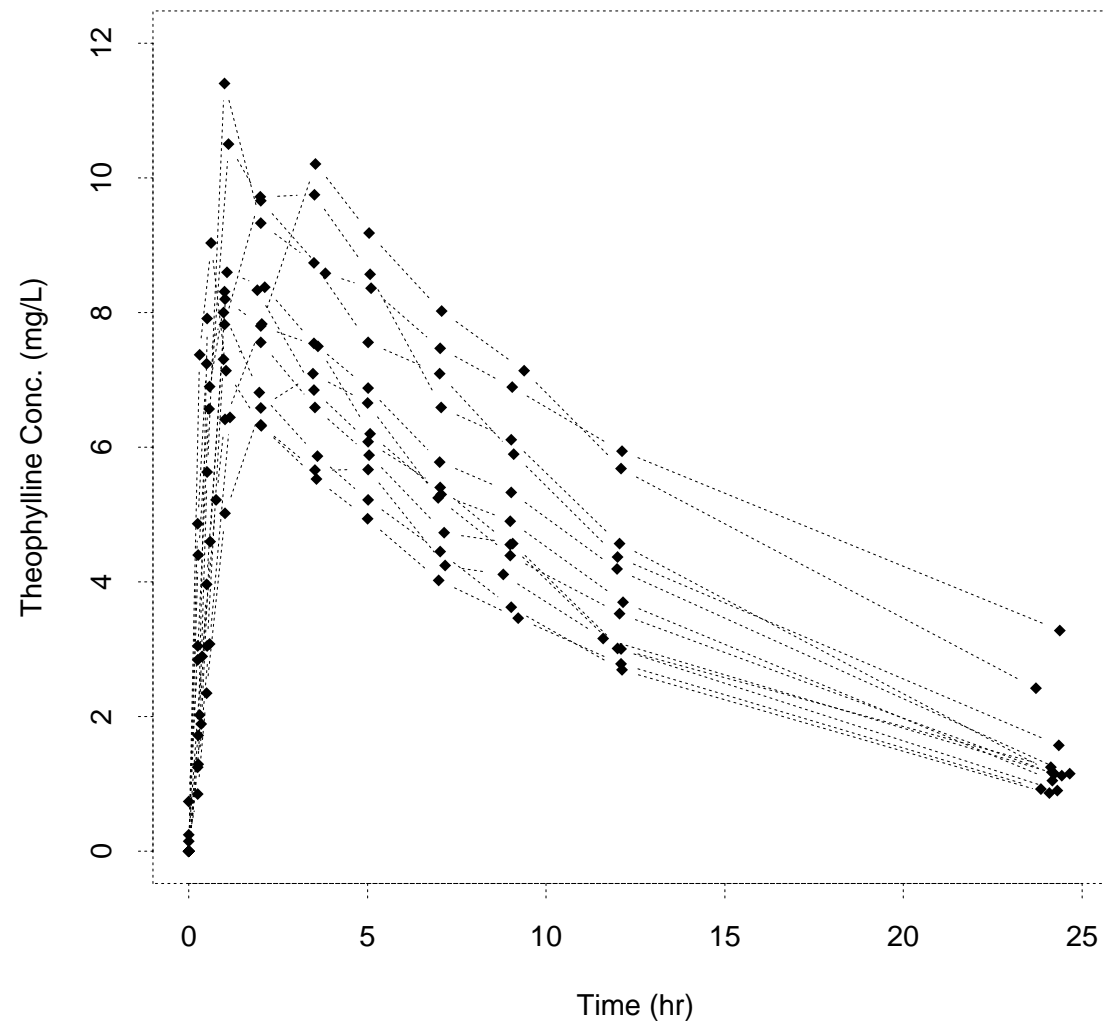
## Pharmacokinetics of theophylline:

- 12 subjects each given *oral dose* at time 0
- *Blood samples* at 10 time points over next 25 hours, *assayed* for *theophylline concentration*

## Questions of interest: *Informally* stated

- Understand processes of *absorption*, *elimination*, *distribution* in the *population* of subjects like these  
⇒ Dosing recommendations
- What is the “*typical*” behavior of these processes?
- To what extent does it *differ* across subjects?

## Data for 12 subjects: Concentration vs. time



**Standard practice:** A “*theoretical model*” for each subject

- Represent the body of  $i$ th subject by a *mathematical compartment model* following oral dose  $D$

**Concentration at time  $t$ :**

$$C_i(t) = \frac{k_{ai}D}{V_i(k_{ai} - k_{ei})} \{ \exp(-k_{ei}t) - \exp(-k_{ai}t) \}$$

- *Fractional absorption rate*  $k_{ai}$ , *fractional elimination rate*  $k_{ei}$ , *volume of distribution*  $V_i$  characterize *absorption*, *elimination*, *distribution* processes for subject  $i$

## Observations:

- *Not balanced* (different times for different subjects)
- Concentration-time patterns *same shape*, but *differ* for different subjects
- *Theory*: This is because  $k_{ai}$ ,  $k_{ei}$ ,  $V_i$  *differ* across subjects  
⇒ Learn about “*typical*” (*average*) *values* and *extent of variation* of  $k_{ai}$ ,  $k_{ei}$ ,  $V_i$  in population of subjects

**Note:** The question of interest needs to refer to the *pharmacokinetic one-compartment model*

**Summary:** Different questions in different settings

- *Characterize* and *compare patterns of change* over time
- Assess *associations* that *evolve over time*
- Learn about features *underlying* observed patterns

**Summary:** Features of data

- *Different* types of response (continuous, discrete)
- Subjects observed only *intermittently*...
- ...at possibly *different* time points with responses we intended to collect *missing* for some subjects (so at the very least not *balanced*)

## 2. Some *ad hoc* approaches

**Dental study:** 16 boys, 11 girls, *distance* measured at 8, 10, 12, 14 years of age, no *missing* observations

- *Focus:* Is dental distance over time (pattern) *different* for boys and girls?

### **Favorite *ad hoc* analysis:**

- *Cross-sectional* analysis comparing *means* (boys vs. girls) at each age 8, 10, 12, 14 (*two-sample t-tests*)
- *P-values:* 0.08, 0.06, **0.01**, **0.001**
- *Conclusion?* *Multiple comparisons?*
- How to “*put this together*” to say something about the *differences in patterns* and *how* they differ? *What are the patterns, anyway?*

**Problem:** We're trying to *force* a familiar analysis to address questions it's not designed to answer!

- *In fact*, what if the data weren't *balanced*?
- Need to start with a formal *model* for the situation that acknowledges the data structure...

**Better:** (*Univariate*) *repeated measures analysis of variance* model

- For subject  $i$  in group  $\ell$  at the  $j$ th time

$$y_{ilj} = \mu + \tau_{\ell} + \gamma_j + (\tau\gamma)_{\ell j} + b_{il} + e_{ilj}, \quad b_{il} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_b^2), \quad e_{ilj} \stackrel{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$$

*Population mean* for group  $\ell$ , time  $j = \mu + \tau_{\ell} + \gamma_j + (\tau\gamma)_{\ell j}$

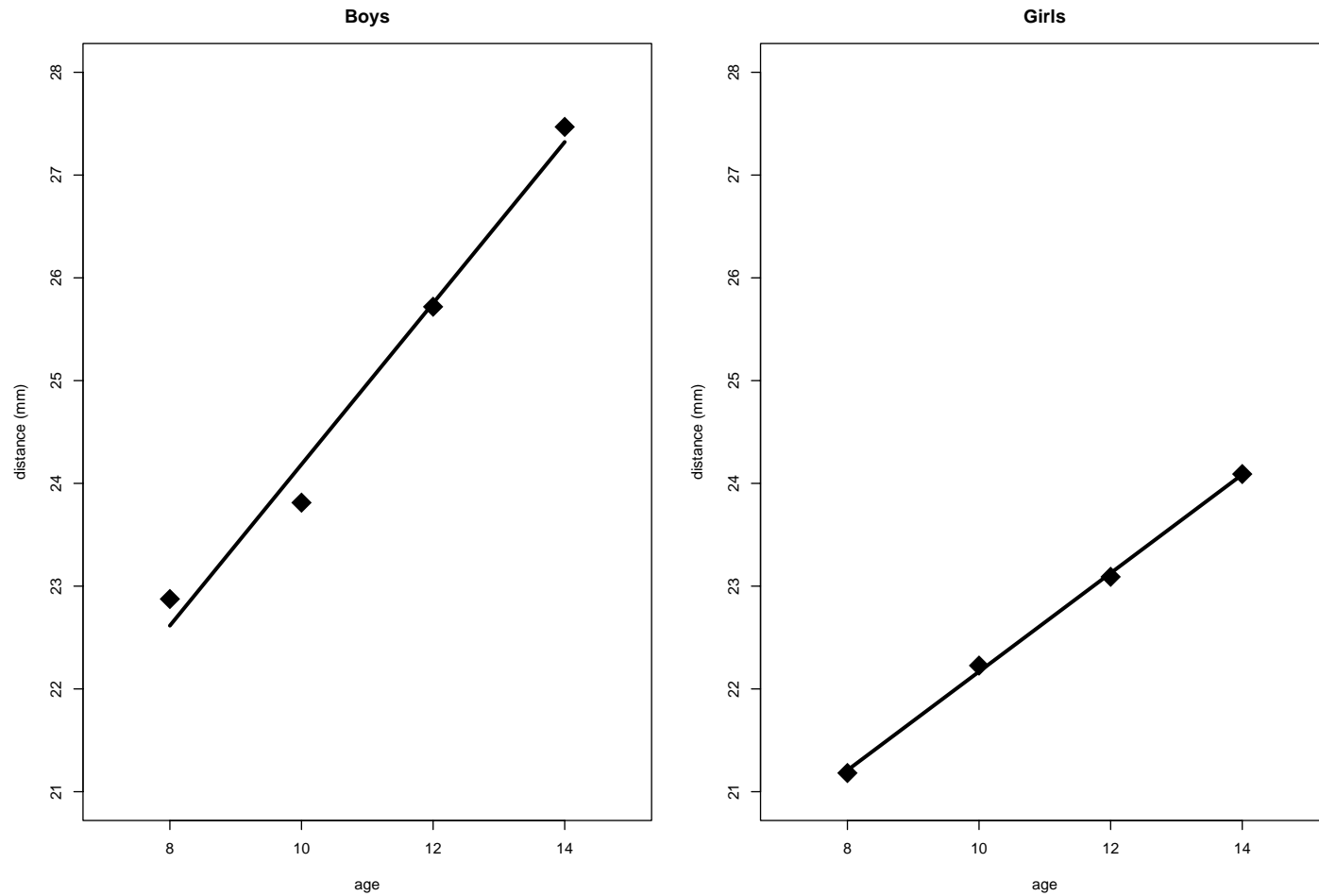
- Is pattern different for girls and boys?  
 $\Rightarrow (\tau\gamma)_{\ell j} = 0$  for all  $\ell, j \Leftrightarrow$  mean profiles are *parallel* across groups

## Drawbacks:

- Requires data to be *balanced*
- May be *too simple* to capture key features of longitudinal data
- Doesn't explicitly acknowledge *time* or exploit apparent smooth, meaningful *patterns*
- What if the data are *discrete*?

**For the dental data:** *Individual child* and *gender-averaged* trajectories look like *straight lines*...

## Gender-averaged trajectories: Means across boys, girls at each time



**Impression:** Population mean distances lie approximately on a *straight line* over time for each gender

**Question of interest, more formally:** *Assuming* that *population means* follow a *straight line pattern over time* for each gender

- Is pattern different for girls and boys?  
⇒ Are the slopes of the *population mean* profiles *different* for boys and girls?

**Perspective:** These are questions about how the *population means* are related over time

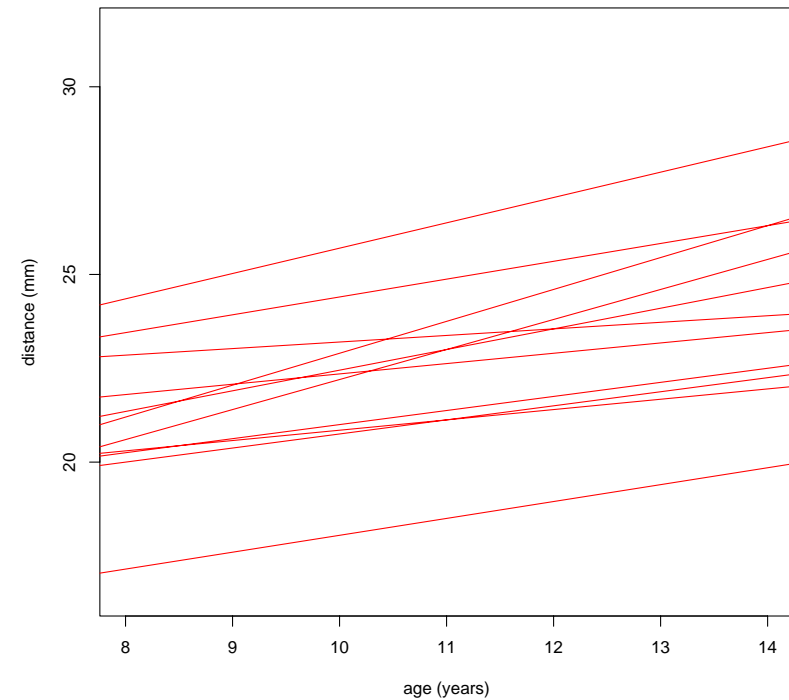
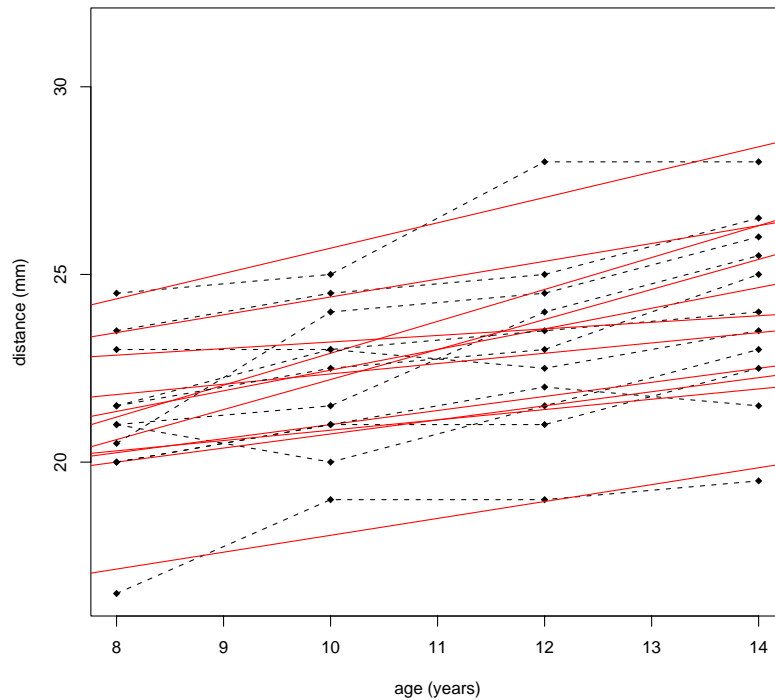
**Suggests:** *Ad hoc* analysis based on *regression model* for *populations* of girls and boys

- For subject  $i$  at age  $t_{ij}$ ,

$$y_{ij} = \beta_{0G} + \beta_{1G}t_{ij} + e_{ij} \text{ if } i \text{ is girl, } y_{ij} = \beta_{0B} + \beta_{1B}t_{ij} + e_{ij} \text{ if } i \text{ is boy,}$$

- Fit by usual OLS, test if  $\beta_{1G} = \beta_{1B}$
- But are  $y_{ij}$  ( $e_{ij}$ ) all *uncorrelated* (required for OLS)?

## Individual trajectories: Girls



**Impression:** *Each girl's* distance measurements follow an approximate *straight line* trajectory with *possibly different slopes* across girls (similarly for boys)

**Question of interest, more formally:** *Assuming* that each child has his/her *own* underlying straight-line trajectory

- Is pattern different for girls and boys?  
⇒ Is the “*typical*” (average) slope among girls *different* from that for boys?

**Perspective:** These are questions about *individual profiles* over time

**Suggests:** *Ad hoc* analysis based on *individual regression models*

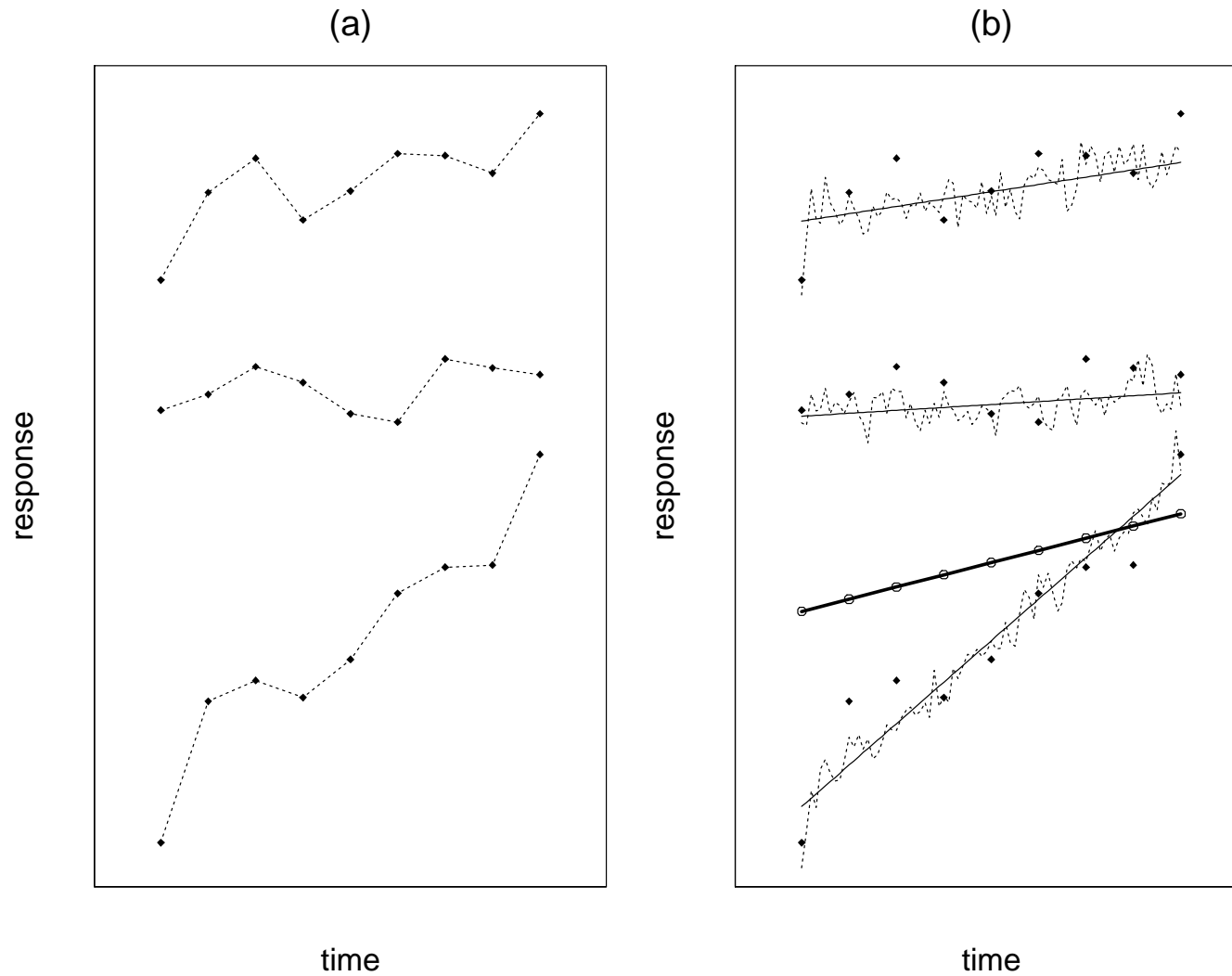
- For subject  $i$  at age  $t_{ij}$ ,  $y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + e_{ij}$ ,  $E(e_{ij}) = 0$
- Fit to *each child* by OLS and do two-sample t-test using estimated individual slopes
- But are  $y_{ij}$  all *uncorrelated*? What if data aren't *balanced*?

**Need a more formal approach. . .**

### 3. How do longitudinal data happen?

**Idea:** *Conceptualize* how longitudinal data come about and use as a basis for developing *formal statistical models* that lead to *appropriate methods* for analysis. . .

**Three hypothetical subjects:** (a) What we see, and (b) a conceptualization of what's underlying it



## Features of the conceptual model:

- Each subject has an “*inherent trend*” or “*trajectory*”
- Actual values might “*fluctuate*” about the trend
- *Errors in measurement* in ascertaining values might occur (continuous response)
- Averaging over all possible values/measurements for all possible subjects in the population at each time yields the **bold** *population mean* profile

## Remarks:

- Individual trajectories and population mean profile *need not* be straight lines (think of theophylline)
- Can think similarly for *discrete* data (Six Cities)

## A key feature: *Correlation*

- Reasonable to suppose that measurements on *different subjects* are *unrelated*  $\Rightarrow$  *independent*, however...
- Measurements on the *same subject* tend to be “*high*” or “*low*” together, so that measurements on the *same subject* are “*more alike*” than measurements from *different subjects*  
 $\Rightarrow$  Measurements on the same subject are *correlated* due to *among-individual* variation
- Values *close together in time* might tend to “*fluctuate*” *similarly*, so that measurements on a given subject are “*more alike*” the *closer together* they are in time  
 $\Rightarrow$  Measurements on the same subject are *correlated* due to *within-individual* covariation

**Result:** A *statistical model* must acknowledge that

- While observations on different subjects may be reasonably thought of as *independent*. . .
- . . . Observations on the same subject are *correlated* due to at least one of these phenomena

**Critical point:** If we *ignore* correlation, we act as though we have *more information* than we actually do, and analyses will be *flawed*

- Can be shown formally by *statistical theory*
- *Statistical models and methods must this acknowledge correlation!*

## 4. Statistical models for longitudinal data

**Two popular types:** Corresponding to the two perspectives on the dental data

- *Subject-specific* models
- *Population-averaged* (aka *marginal*) models
- Depending on the *questions* in a particular situation, one may be more suitable than the other

**Here:** In terms of dental data (*continuous* response, *straight-line* population mean and individual patterns) and then generalize

## Subject-specific model:

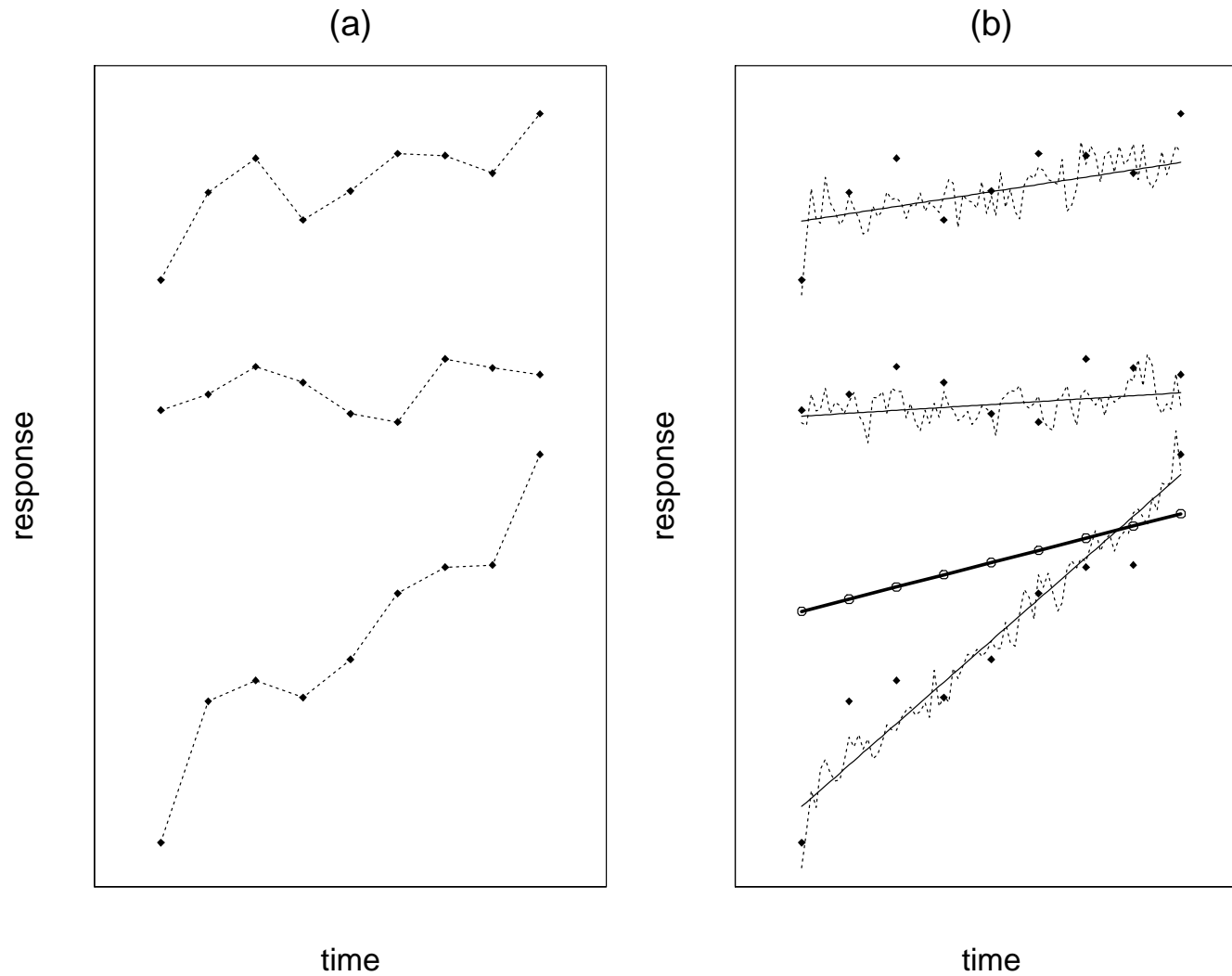
- Model *individual behavior*
- Questions of interest are about “*typical*” (*average*) such behavior

**Dental data:** For *randomly chosen* subject  $i$ , measure  $y_{ij}$  at several times  $t_{ij}$  (need not be the same for all  $i$ )

$$y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + \underbrace{e_{f,ij} + e_{me,ij}}_{e_{ij}}$$

- $\beta_{0i}, \beta_{1i}$  are *individual-specific* intercept, slope dictating  $i$ 's “*inherent trajectory*”  $\beta_{0i} + \beta_{1i}t_{ij}$
- $e_{f,ij}$  is *mean-zero* deviation from inherent trajectory due to “*fluctuation*” at  $t_{ij}$
- $e_{me,ij}$  is *mean-zero* deviation from inherent trajectory due to *error in measurement* at  $t_{ij}$

**Conceptualization:**  $y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + \underbrace{e_{f,ij} + e_{me,ij}}_{e_{ij}}$



$$y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + \underbrace{e_{f,ij} + e_{me,ij}}_{e_{ij}}$$

- $e_{f,ij}$  for times close together might be in *same direction*  
 $\Rightarrow$  *within-subject (auto)correlation* across  $j$
- $e_{me,ij}$  are likely *independent* across  $j$  (a new error each time)
- $\beta_{0i}, \beta_{1i}$  come from a *population* of intercepts, slopes

$$\beta_{0i} = \gamma_{0G} + b_{0i}, \quad \beta_{1i} = \gamma_{1G} + b_{1i} \quad \text{if } i \text{ is a girl}$$

$$\beta_{0i} = \gamma_{0B} + b_{0i}, \quad \beta_{1i} = \gamma_{1B} + b_{1i} \quad \text{if } i \text{ is a boy}$$

$b_{0i}, b_{1i}$  are *mean-zero random effects* describing how  $i$  deviates from the “*typical*” (*mean*) intercept and slope

- $y_{i1}, \dots, y_{i4}$  *all depend* on  $b_{0i}, b_{1i} \Rightarrow$  *among-subject correlation*

$$y_{ij} = \beta_{0i} + \beta_{1i}t_{ij} + \underbrace{e_{f,ij} + e_{me,ij}}_{e_{ij}}, \quad \begin{aligned} \beta_{0i} &= \gamma_{0G \text{ or } B} + b_{0i} \\ \beta_{1i} &= \gamma_{1G \text{ or } B} + b_{1i} \end{aligned}$$

### Technically speaking:

- *Within-subject autocorrelation:*  $(e_{f,i1}, \dots, e_{f,i4})^T$  is *multivariate normal* with *covariance matrix*  $\sigma_f^2 H_i$
- *Measurement error:*  $(e_{me,i1}, \dots, e_{me,i4})^T$  is *multivariate normal* with *diagonal covariance matrix*  $\sigma_e^2 I_i$
- “Steep/shallow” slopes associated with “high/low” intercepts  
 $\Rightarrow (b_{0i}, b_{1i})^T$  correlated with *covariance matrix*  $D$  (assumed *multivariate normal*)

**Combining:** 
$$y_{ij} = \gamma_{0G \text{ or } B} + \gamma_{1G \text{ or } B}t_{ij} + b_{0i} + b_{1i}t_{ij} + e_{ij}$$

**Can summarize in matrix form...** 
$$y_i = (y_{i1}, \dots, y_{i4})^T$$

## Linear mixed effects model:

$$y_i = X_i\gamma + Z_ib_i + e_i$$

$$\gamma = \begin{pmatrix} \gamma_{0G} \\ \gamma_{1G} \\ \gamma_{0B} \\ \gamma_{1B} \end{pmatrix}, \quad b_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}, \quad Z_i = \begin{pmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{i4} \end{pmatrix}$$

$$X_i = \begin{pmatrix} 1 & t_{i1} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{i4} & 0 & 0 \end{pmatrix} \text{ for } i \text{ a girl, } \quad X_i = \begin{pmatrix} 0 & 0 & 1 & t_{i1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & t_{i4} \end{pmatrix} \text{ for } i \text{ a boy}$$

$$E(y_i) = X_i\gamma, \quad \text{var}(y_i) = Z_iDZ_i^T + \sigma_f^2H_i + \sigma_e^2I_i = V_i$$

so that

$$y_i \sim \mathcal{MVN}(X_i\gamma, V_i)$$

## Features:

- Questions about “*typical*” *individual behavior* are questions about  $\gamma$
- The *covariance matrix*  $V_i$  has a specific form with *separate components* for each type of correlation, which the analyst can specify
- *No requirement* for balance

## For the dental data:

- “*Fluctuations*” in physical distance unlikely? Even if so, 2 years is a long time, suggests  $H_i = I_i$

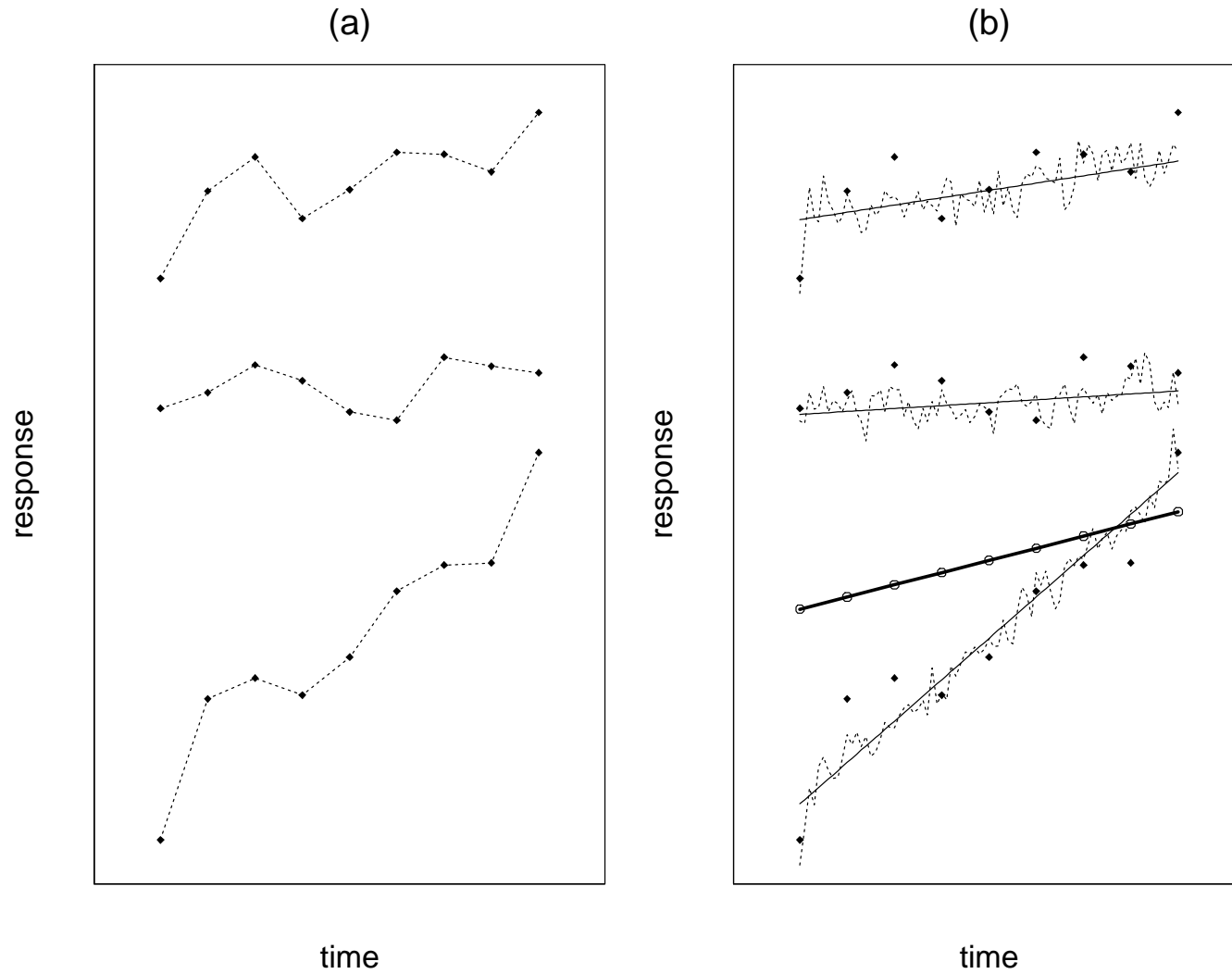
$$\Rightarrow V_i = Z_i D Z_i^T + \underbrace{(\sigma_f^2 + \sigma_e^2)}_{\sigma^2} I_i, \quad y_i \sim \mathcal{MVN}(X_i \gamma, V_i)$$

- Is “*typical*” slope for girls *different* from that for boys?  
 $\Rightarrow$  Test  $\gamma_{1G} = \gamma_{1B}$

## Population-averaged model:

- Model *population behavior* by *directly modeling* the *population mean profile*; i.e.,  $E(y_i)$
- Questions are about how *population means* are related over time
- Instead of worrying about separate components of  $\text{var}(y_i)$  (within- and among-individual sources of correlation), just model their *combined effect* directly

# Conceptualization:



**Dental data:** For randomly chosen subject  $i$  measure  $y_{ij}$  at several times  $t_{ij}$  (need not be the same for all  $i$ )

$$y_{ij} = \beta_{0G} + \beta_{1G}t_{ij} + \epsilon_{ij} \text{ for girls, } y_{ij} = \beta_{0B} + \beta_{1B}t_{ij} + \epsilon_{ij} \text{ for boys}$$

- E.g.,  $\beta_{0G} + \beta_{1G}t_{ij}$  is the **bold** *population mean profile* for girls
- $\epsilon_{ij}$  is a deviation from the population mean due to the *sum total* of among-subject variation, and within-subject fluctuation and measurement error at  $t_{ij}$

Thus,  $\epsilon_{ij}$  are *correlated*  $\Rightarrow$  specify a *covariance matrix*

- Question of whether the patterns are different for boys and girls:  
Are the *slopes of the population mean profiles* the same?  
 $\Rightarrow$  Test  $\beta_{1G} = \beta_{1B}$

**Population-averaged model:** In matrix form

$$y_i = X_i\beta + \epsilon_i, \quad \beta = \begin{pmatrix} \beta_{0G} \\ \beta_{1G} \\ \beta_{0B} \\ \beta_{1B} \end{pmatrix}$$

$$X_i = \begin{pmatrix} 1 & t_{i1} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{i4} & 0 & 0 \end{pmatrix} \text{ for } i \text{ a girl, } X_i = \begin{pmatrix} 0 & 0 & 1 & t_{i1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & t_{i4} \end{pmatrix} \text{ for } i \text{ a boy}$$

- Implies  $E(y_i) = X_i\beta$  (average over all individuals, fluctuations, errors)
- Choose a “*working model*” for the covariance matrix  $\text{var}(\epsilon_i) = \Sigma_i$  that (hopefully) captures the overall combined correlation
- Thus, model is  $E(y_i) = X_i\beta$ ,  $\text{var}(y_i) = \Sigma_i$  (can add normality)

## Contrasting the models:

**Subject-specific:**  $y_i = X_i\gamma + Z_ib_i + e_i$ , and *averaging* these *individual models* over all  $b_i$  and  $e_i$  (so over all subjects, fluctuations, and measurement errors) gives  $E(y_i) = X_i\gamma$

**Population-averaged:**  $y_i = X_i\beta + \epsilon_i$ , and we model this average *directly* as  $E(y_i) = X_i\beta$

**Result:** The models for the population means at all time points are of the *same form* with either model!

- Thus  $\gamma$  and  $\beta$  describe the *same thing* (population mean profile), so are really the same ...
- ...and we can interpret them either way, e.g., “*typical slope*” or *slope of the population average profile*!
- The distinction between *subject-specific* and *population-averaged* ends up not mattering, so choose the interpretation you like best!

**Warning:** This changes when the model is *nonlinear*...

## What about discrete data? Six Cities data

**Subject-specific model:** Model *individual propensity* for respiratory infection ( $y_{ij} = 1$ ) when exposed to maternal smoking  $x_{ij}$  (=0 or 1)

$$\log \left( \frac{P(y_{ij} = 1|b_i)}{1 - P(y_{ij} = 1|b_i)} \right) = \beta_{0i} + \beta_{1i}x_{ij}, \quad \beta_{0i} = \gamma_0 + b_{0i}, \quad \beta_{1i} = \gamma_1 + b_{1i}$$

$P(y_{ij} = 1|b_i)$  is the probability of infection for *child  $i$  in particular*

- $\beta_{0i}$  is the log odds of respiratory infection *for child  $i$*  when mother does not smoke  
 $\Rightarrow \gamma_0$  is the “*typical*” (mean) value of the log odds for children in the population
- $\beta_{1i}$  is the log odds ratio for respiratory infection when *child  $i$*  is exposed to smoking relative to not, and  $\gamma_1$  is the “*typical*” value of the log odds ratio  
 $\Rightarrow$  Thus,  $\gamma_1 = 0$  asks whether the “*typical*” odds ratio for children in the population is equal to 1

**Population-averaged model:** Model “*average propensity*” for respiratory infection in *the population directly*

$$\log \left( \frac{P(y_{ij} = 1)}{1 - P(y_{ij} = 1)} \right) = \beta_0 + \beta_1 x_{ij}$$

$P(y_{ij} = 1)$  is the probability a child in the population under maternal smoking  $x_{ij}$  will have a respiratory infection

- $\beta_0$  is the log odds of respiratory infection in the population of children whose mothers don't smoke
- $\beta_1$  is the log odds ratio for respiratory infection if the population were exposed to smoking relative to not
- Thus,  $\beta_0$  and  $\beta_1$  describe what happens “*on average*” in the population (as opposed to for a particular individual child)
- ... and  $\beta_1 = 0$  asks whether the odds ratio *for the population* is equal to 1

## Contrasting the models:

**Population-averaged:** 
$$P(y_{ij} = 1) = \frac{\exp(\beta_0 + \beta_1 x_{ij})}{1 + \exp(\beta_0 + \beta_1 x_{ij})}$$

- A model for the *average* over all children in the population

**Subject-specific:** 
$$P(y_{ij} = 1 | b_i) = \frac{\exp(\gamma_0 + \gamma_1 x_{ij} + b_{0i} + b_{1i} x_{ij})}{1 + \exp(\gamma_0 + \gamma_1 x_{ij} + b_{0i} + b_{1i} x_{ij})}$$

- A model *specifically* for the  $i$ th child
- The *average* of this over all children (so over all  $b_{0i}, b_{1i}$ ) is a *very complicated* mess, that *does not* have the same form as the population-averaged model above!

**Result:** In contrast to linear models, for *nonlinear models* like this,  $\beta$  and  $\gamma$  have distinct interpretations

**Back to the examples:** Which perspective/model makes more sense?

- *Dental data*: linear model, can go either way! (Either interpretation valid!)
- *Six Cities data*: For inferences to be used for *public policy* recommendations, what happens *on average* in the population is usually more relevant than what happens to individuals  
 $\Rightarrow$  population-averaged model
- *Theophylline PK data*: Interest clearly focused on “*typical values*” and variation of  $k_{ai}$ ,  $k_{ei}$ ,  $V_i \Rightarrow$  subject-specific model

$$y_{ij} = \frac{k_{ai}D}{V_i(k_{ai} - k_{ei})} \{e^{-k_{ei}t} - e^{-k_{ai}t}\} + e_{ij}$$

$$k_{ai} = \gamma_1 + b_{k_a,i}, \quad k_{ei} = \gamma_2 + b_{k_e,i}, \quad V_i = \gamma_3 + b_{V,i}$$

$\Rightarrow$  Subject-specific model

*Nonlinear mixed effects model*

**Message:** Choose the model that best addresses the questions!

## 5. Methods for implementation

**Linear models:** *Covariance matrix* plays key role!

**Population-averaged models:** Solve *generalized estimating equations* (GEE) for  $\beta$  and parameters in  $\text{var}(y_i) = \Sigma_i$

$$\sum_{i=1}^n X_i^T \hat{\Sigma}_i^{-1} (y_i - X_i \hat{\beta}) = 0 \quad \text{which yields} \quad \hat{\beta} = \left( \sum_{i=1}^n X_i^T \hat{\Sigma}_i^{-1} X_i \right)^{-1} \sum_{i=1}^n X_i^T \hat{\Sigma}_i^{-1} y_i$$

- SAS proc genmod

**Subject-specific models:** *Likelihood* methods based on  $y_i \sim \mathcal{MVN}(X_i \gamma, V_i)$  to estimate  $\gamma$  and parameters in  $V_i$  lead to the *same approach!*

$$\hat{\gamma} = \left( \sum_{i=1}^n X_i^T \hat{V}_i^{-1} X_i \right)^{-1} \sum_{i=1}^n X_i^T \hat{V}_i^{-1} y_i$$

- SAS proc mixed, Splus/R lme()

**Nonlinear models:** *Covariance matrix* still plays key role, but SS and PA *no longer the same*

**Population-averaged models:** Solve similar *generalized estimating equations* (GEEs) for  $\beta$  and parameters in  $\text{var}(y_i) = \Sigma_i$

- SAS proc genmod

**Subject-specific models:** Much messier! (Likelihood methods)

- E.g., for binary data, must find the average of

$$P(y_{ij} = 1|b_i) = \frac{\exp(\gamma_0 + \gamma_1 x_{ij} + b_{0i} + b_{1i} x_{ij})}{1 + \exp(\gamma_0 + \gamma_1 x_{ij} + b_{0i} + b_{1i} x_{ij})}$$

over  $b_{0i}, b_{1i}$ , an *integral*

- *Generalized linear mixed* or *nonlinear mixed effects* model
- SAS proc nlmixed, %glimmix, %nlinmix, Splus/R nlme()

## 6. Discussion

**Main message:** Specialized *statistical models* are required for longitudinal data analysis – choose the one that's right for you!

**Benefit:** Understanding the *basis* for the models and the *role of correlation* is essential to understanding how to use the software!

## What we didn't talk about: Lots!

- More *advanced* modeling considerations
- How to choose appropriate *covariance models* and what happens if we're *wrong*
- How to *select* the best model and *diagnose* how well a model fits
- Details of *implementation*
- What happens if *assumptions* are incorrect
- How to handle *missing data* and *dropout*
- Other types of models (e.g., *transition* models)

**Where to learn more:** Some references

- Diggle, P.J., Heagerty, P., Liang, K.-Y., and Zeger, S.L. (2002) *Analysis of Longitudinal Data, 2nd Edition*, Oxford University Press.
- Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*, Springer.
- Davidian, M. and Giltinan, D.M. (1995) *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall/CRC Press.

**Where to get a copy of these slides (and more):**

<http://www.stat.ncsu.edu/~davidian>