An Introduction to Modeling and Analysis of Longitudinal Data

Marie Davidian
Department of Statistics
North Carolina State University

http://www.stat.ncsu.edu/∼davidian

(*a copy of these slides is available at this website*)

Outline

1. Some examples and questions of interest

Longitudinal studies: Studies where a *response* is observed on each subject/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* that acknowledge these features and the questions of interest are needed, which lead to appropriate *methods*
- Understanding the *models* is *critical* to using the *software*

1. Some examples and questions of interest

First, an "ideal" situation. . .

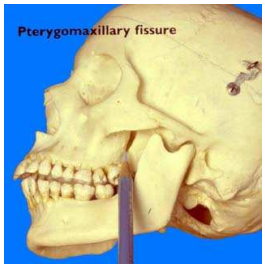
"World-famous" dental study: Pothoff and Roy (1964)

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

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*?

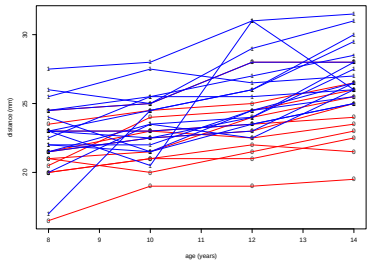
1. Some examples and questions of interest



From web pages by Professor John B. Ludlow, UNC-Chapel Hill School of Dentistry

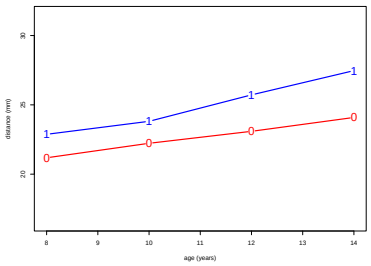
1. Some examples and questions of interest

All data ("**spaghetti plot**"): 0 = *girl*, 1 = *boy*



1. Some examples and questions of interest

Sample mean dental distances: Sample averages across all *boys* and all *girls* at each age



1. Some examples and questions of interest

Observations:

- *All children* have all 4 measurements at the *same* time points (ages) ("*balanced*")
- Boys seem to be "*higher*" than girls overall
- 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 *mean* of distance (across boys and across girls) follows an approximate *straight line* pattern (with some "*jitter*")

1. Some examples and questions of interest

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)
- Data for three children: city, age, smoking, respiratory status

Portage	9	1	1	10	1	0	11	1	0	12	1	0
Kingston	9	0	0	10	0	0	11	0	0	12	0	0
Portage	9	0	0	10	.	.	11	.	.	12	.	.

- *Discrete* (*binary*) response
- *Missing* data at some ages for some mother-child pairs (*balance*?)

1. Some examples and questions of interest

Questions of interest: *Informally* stated

- Is there an *association* between risk of respiratory infection and mother's smoking behavior?
- Does the risk of respiratory infection *change with age*? Does the association *change with age*?

Observations:

- Graphical depiction not really informative (*binary* response)
- *Crude* summary for *Portage* at each age

Age	Prop. Mom Smoke	Prop. with RS=1
9	0.62	0.38
10	0.64	0.37
11	0.69	0.23
12	0.75	0.08

1. Some examples and questions of interest

Pharmacokinetics of theophylline:

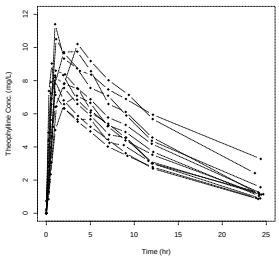
- 12 subjects each given *oral dose* at time 0
- *Blood samples* at 10 time points over next 25 hours, *assayed* for *theophylline concentration*

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 *vary* across subjects?
- Is some of this variation associated with *subject characteristics*?

1. Some examples and questions of interest

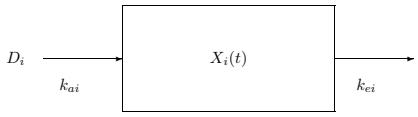
Data for 12 subjects: Concentration vs. time



1. Some examples and questions of interest

Standard practice: A "*theoretical model*" for each subject

- Represent the body of ith subject by a *mathematical compartment model*
- *One compartment model with first-order absorption and elimination* following oral dose D_i



- $X_i(t)$ = *amount* of drug in blood at time t
  V_i = hypothetical "*volume*" of the blood compartment

1. Some examples and questions of interest

Concentration at time t: Solve the corresponding *differential equations* for $X_i(t)$, divide by volume

$$C_i(t) = \frac{k_{ai}D_i}{V_i(k_{ai}-k_{ei})}\{\exp(-k_{ei}t) - \exp(-k_{ai}t)\}, \quad k_{ei} = Cl_i/V_i$$

- *Fractional absorption rate* k_{ai} characterizes the *absorption* process for subject i
- *Clearance rate* Cl_i characterizes the *elimination* process for subject i
- *Volume of distribution* V_i characterizes the process of *distribution* in the body for subject i
- For subject i, we observe $C_i(t)$ at several time points subject to some *variation* (more later...)

1. Some examples and questions of interest

Observations:

- *Not balanced* (different times for different subjects)
- Concentration-time patterns have *same general shape* but *differ* for different subjects
- *Theory:* This is because k_{ai}, Cl_i, V_i *differ* across subjects
  $\Rightarrow$ Learn about "*typical*" (*average*) *values* and *extent of variation* of k_{ai}, Cl_i, V_i in the population of subjects
- *Furthermore*, part of this *variation* in k_{ai}, Cl_i, V_i across subjects might be systematically *associated* with *weight*, *renal function*, etc, and we'd like to know about this!

Note: The questions of interest need to refer to the *pharmacokinetic one-compartment model*

1. Some examples and questions of interest

Summary: Different questions in different settings

- *Characterize* and *compare patterns of change* in response over time
- Assess *associations* between response and other factors that *evolve over time*
- Learn about meaningful features *underlying* observed patterns and how they *vary* in the population of subjects

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. First steps

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

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

Favorite ad hoc analysis of my clinician friends:

- *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*?

2. First steps

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 *statistical model* for the situation that acknowledges the data structure...

Statistical model:

- *Informally* – a description of the *mechanisms* by which data are thought to arise
- *More formally* – a probability distribution that describes how observations we see take on their values
- In order to talk about *analysis*, we need to first identify an appropriate *statistical model*...

2. First steps

One traditional, popular statistical model: (*Univariate*) *repeated measures analysis of variance* model (*continuous* response)

- $Y_{i\ell j}$ = *response* for subject i in group ℓ at jth time
- The model says

$$Y_{i\ell j} = \mu + \tau_\ell + \gamma_j + (\tau\gamma)_{\ell j} + b_{i\ell} + e_{i\ell j}, \quad b_{i\ell} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_b^2), \quad e_{i\ell j} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2)$$

- *Population mean* for group ℓ, time j is $\mu + \tau_\ell + \gamma_j + (\tau\gamma)_{\ell j}$
- $b_{i\ell}$ allows responses for subject i in group ℓ to be "*high*" or *low*" relative to the mean for the group (by *same amount* at all times)
- $e_{i\ell j}$ allows responses for subject i in group ℓ *furthermore* to vary because of things like *measurement error*
- Is the *pattern of mean change* different for girls and boys?
  $\Rightarrow (\tau\gamma)_{\ell j} = 0$ for all $\ell, j \Leftrightarrow$ mean profiles are *parallel* across groups

2. First steps

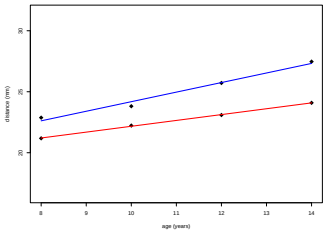
Some drawbacks: So what's wrong with this model?

- Requires data to be *balanced* (*same* j for all subjects)
- Allows the *population mean* for each group at each time to be *anything* – no *relationship* of means over each time
- So doesn't explicitly acknowledge apparent smooth, meaningful *patterns* over time
- Doesn't explicitly acknowledge *time itself*
- May be *too simple* to capture the true pattern of *variation* in longitudinal data
- What if the response is *discrete*?

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

2. First steps

Gender-averaged trajectories: *Sample means* across boys, girls at each time *straight lines* superimposed



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

2. First steps

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

- Is the *pattern of mean change* different for girls and boys?
  $\Rightarrow$ Are the *slopes* of the *population mean* profiles *different* for boys and girls?

Perspective: This is a question about the *population means* and how they are *related* over time

- Need a *statistical model* that incorporates our belief that they lie on a *straight line* for each gender

2. First steps

Suggests: A first shot at a *statistical model*

- Y_{ij} = distance for child $i = 1, \ldots, 27$ at time $t_{ij} = 8, 10, 12, 14$
- G_i = *gender indicator* = 0 if i is a girl, = 1 if i is a boy
- *Observed data for each child*: $(Y_{i1}, \ldots, Y_{i4}, G_i)$ $(j = 1, 2, 3, 4)$
- *Assume* the *population means* lie on a *straight line* for each gender
- For subject i at age t_{ij}

$Y_{ij} = \beta_{0G} + \beta_{1G} t_{ij} + \epsilon_{ij}$ if i is girl, $Y_{ij} = \beta_{0B} + \beta_{1B} t_{ij} + \epsilon_{ij}$ if i is boy

$$Y_{ij} = \beta_{0G}(1 - G_i) + \beta_{0B} G_i + \beta_{1G}(1 - G_i) + \beta_{1B} G_i + \epsilon_{ij}$$

- ϵ_{ij} is a *mean-zero* "*deviation*" that accounts for fact that the distance we observe at t_{ij} for i is not *exactly equal* to

Population mean for girls at t_{ij} = $\beta_{0G} + \beta_{1G} t_{ij}$
Population mean for boys at t_{ij} = $\beta_{0B} + \beta_{1B} t_{ij}$

2. First steps

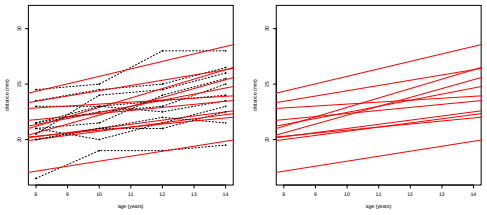
This looks great! So how to we do an *analysis* based on this model to answer the question?

- Fit by usual OLS, test if $\beta_{1G} = \beta_{1B}$?
- But are Y_{ij} (ϵ_{ij}) all *uncorrelated* or *independent* (required for usual OLS analysis)?
- More coming shortly...

We can take another perspective...

2. First steps

Individual trajectories: Girls



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

2. First steps

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: *Another* statistical model

- For child i at age t_{ij},

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}$$

- β_{0i}, β_{1i} are the *child-specific* intercept and *slope* for i's assumed straight line
- e_{ij} is a *mean-zero* "*deviation*" accounting for the "*jitter*" in i's distances about his/her *child-specific* line

2. First steps

Statistical model, continued: $Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}$

- Each child has his/her *own* (β_{0i}, β_{1i})
- These *vary* across children in each gender group
  ⇒ the (β_{0i}, β_{1i}) come from a *probability distribution* that describes this variation (more coming up...)

Analysis regarding different patterns of change?

- *The question becomes*: Is the *mean* of β_{1i} values in the *population of girls* equal to that for the *population of boys*?
- *Ad hoc approach*: Fit to *each child* by OLS and do two-sample t-test using the *estimated* individual slopes as the "*data*"
- But the *estimated* slopes are *not* the *true* slopes!!

2. First steps

Need to think more carefully and adopt a more formal approach...

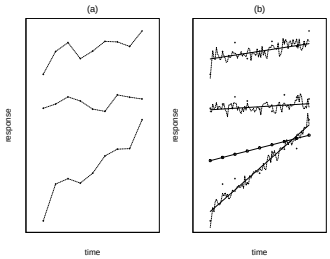
3. How do longitudinal data happen?

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

Consider continuous response...

3. How do longitudinal data happen?

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



3. How do longitudinal data happen?

Features of the conceptual model: Think about the *dental data*

- Each subject has an "*inherent trend*" or "*trajectory*" describing the overall "*track*" s/he follows over the *longer term*
- Actual values of the response might "*fluctuate*" about the trend
- *Errors in measurement* in ascertaining values might occur (continuous response)
- *Averaging* over all trajectories, fluctuations, measurement errors for all possible subjects in the population at each time yields the **bold** *population mean* profile

3. How do longitudinal data happen?

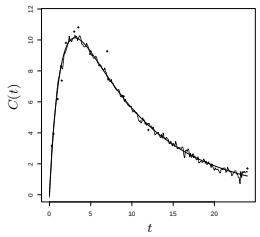
In the picture: $Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + e_{ij}$

- The *individual-specific* (β_{0i}, β_{1i}) determine the "*inherent trajectory*" over the long haul...
- ...and determine at *any time* where i's trajectory sits relative to the **population mean** profile
- The *combined effects* of shorter term "*fluctuations*" and *measurement error* produce the responses we *actually observe*

3. How do longitudinal data happen?

Remarks:

- Individual trajectories and population mean profile *need not* be straight lines (think of theophylline)



3. How do longitudinal data happen?

Remarks:

- Can think similarly for *discrete* data (Six Cities)
- *Usual assumption*: There is no *measurement error* associated with *binary* or *count* responses

3. How do longitudinal data happen?

A key feature: *Correlation*

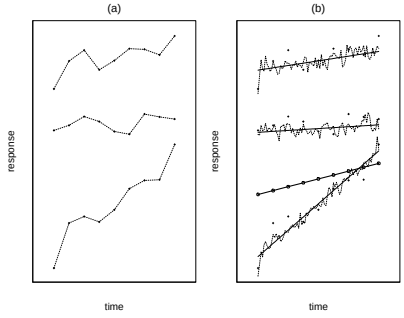
- Reasonable to suppose that responses from *different subjects* are *unrelated* ⇒ *independent*, however...
- Responses from the *same subject* tend to be "*alike*" because they follow the *same underlying trajectory* (e.g., may be "*high*" or "*low*" together as in the dental data)

  ⇒ Responses from the same subject are *correlated* due to *among-individual* variation (heterogeneity)
- Values *close together in time* might tend to "*fluctuate*" *similarly*, so that responses from a given subject are "*more alike*" the *closer together* they are in time

  ⇒ Measurements on the same subject are *correlated* due to *within-individual* covariation

3. How do longitudinal data happen?

Conceptualization:



3. How do longitudinal data happen?

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 and *pretend* all Y_{ij} are *uncorrelated* or *independent*, we *misrepresent* the amount of *information* we actually have, 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

- *Population-averaged* models
- *Subject-specific* models
- Depending on the *questions* in a particular situation, one may be more suitable then the other

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

Take the second perspective first...

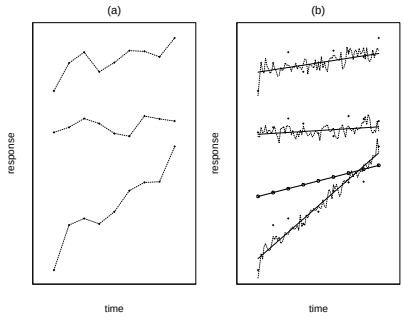
4. Statistical models for longitudinal data

Subject-specific model:

- Model *individual behavior*
- Questions of interest are about "*typical*" (*average* or *mean*) such behavior

4. Statistical models for longitudinal data

Conceptualization:



4. Statistical models for longitudinal data

Conceptualization: For *randomly chosen* subject i, i has an associated *stochastic process*: For any time t

$$Y_i(t) = \beta_{0i} + \beta_{1i}t + \underbrace{e_{f,i}(t) + e_{me,i}(t)}_{e_i(t)}$$

- "*Inherent trajectory*" $\beta_{0i} + \beta_{1i}t$ throughout time *dictated* by i's *subject-specific* intercept β_{0i} and slope β_{1i}
- $e_{f,i}(t)$ is a *mean-zero deviation* due to the *fluctuation* at t
- So $\beta_{0i} + \beta_{1i}t + e_{f,i}(t)$ is the *actual response value* that could be seen at t if there were no *measurement error*
- $e_{me,i}(t)$ is a *mean-zero deviation* due to *measurement error* in ascertaining this value
- $e_i(t)$ is the resulting *overall deviation* (*mean-zero*)

4. Statistical models for longitudinal data

What we observe: A random sample of subjects $i = 1, \ldots, n$, each at *intermittent* times t_{ij}, $j = 1, \ldots, m_i$, say (need not be *the same* for all i)

Thus: We observe $Y_{ij} = Y_i(t_{ij})$, $j = 1, \ldots, m_i$, where

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

- $e_{f,ij} = e_{f,i}(t_{ij})$, $e_{me,ij} = e_{me,i}(t_{ij})$
- With some *assumptions* about β_{0i}, β_{1i}, $e_{f,i}(t)$, and $e_{me,i}(t)$, we have a popular *statistical model*!

4. Statistical models for longitudinal data

Within subjects: $Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \underbrace{e_{f,ij} + e_{me,ij}}_{e_{ij}}$ for subject i

- $e_{f,i}(t)$ and $e_{f,i}(t')$ for times t and t' close together might tend to be in the *same direction* relative to the "*inherent trend*"
  $\Rightarrow$ *within-subject* (*auto*)*correlation*
- We would expect $e_{f,ij}$ close together in time to be *positively correlated*
- *Measuring devices* tend to commit *haphazard* errors
  $\Rightarrow$ $e_{me,i}(t)$ and $e_{me,i}(t')$ might be *unrelated* for any times t and t'
- We would expect $e_{me,ij}$ to all be *mutually independent* across j

4. Statistical models for longitudinal data

Among subjects: (β_{0i}, β_{1i}) are from a *population* of intercepts, slopes

- For *unrelated subjects* drawn at *random*, (β_{0i}, β_{1i}) pairs are *independent* across i
- $\beta_{0i} = \gamma_{0G} + b_{0i}, \quad \beta_{1i} = \gamma_{1G} + b_{1i}$ if i is a girl
  $\beta_{0i} = \gamma_{0B} + b_{0i}, \quad \beta_{1i} = \gamma_{1B} + b_{1i}$ if i is a boy
- b_{0i}, b_{1i} are *mean-zero random effects independent* across i describing how i *deviates* from the "*typical*" (*mean*) intercept (γ_{0G} or γ_{0B}) and slope (γ_{1G} or γ_{1B})
- More succinctly

$$\beta_{0i} = \gamma_{0G}(1 - G_i) + \gamma_{0B} G_i + b_{0i}, \quad \beta_{1i} = \gamma_{1G}(1 - G_i) + \gamma_{1B} G_i + b_{1i}$$

- $Y_{i1}, \ldots, Y_{im_i}$ *all depend* on b_{0i}, b_{1i} $\Rightarrow$ *correlation* due to *among-subject heterogeneity*

4. Statistical models for longitudinal data

Summarizing:

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

$$\beta_{0i} = \gamma_{0G}(1 - G_i) + \gamma_{0B} G_i + b_{0i}, \quad \beta_{1i} = \gamma_{1G}(1 - G_i) + \gamma_{1B} G_i + b_{1i}$$

Remaining: Assumptions on $e_{f,ij}$, $e_{me,ij}$, b_{0i}, b_{1i} that *operationalize* what we've said...

4. Statistical models for longitudinal data

Formally: *Normality* is standard assumption

- *Within-subject autocorrelation:* $(e_{f,i1}, \ldots, e_{f,im_i})^T$ is *multivariate normal* with mean $\mathbf{0}$ and *covariance matrix* $\sigma_f^2 \boldsymbol{H}_i$
- *Measurement error:* $(e_{me,i1}, \ldots, e_{me,im_i})^T$ is *multivariate normal* with mean $\mathbf{0}$ and *diagonal covariance matrix* $\sigma_e^2 \boldsymbol{I}_i$
- So $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{im_i})^T \sim \mathcal{N}(\mathbf{0}, \sigma_f^2 \boldsymbol{H}_i + \sigma_e^2 \boldsymbol{I}_i)$
- "Steep/shallow" slopes associated with "high/low" intercepts
  $\Rightarrow$ $(b_{0i}, b_{1i})^T$ are *correlated* with mean $\mathbf{0}$ and *covariance matrix* $\boldsymbol{D}$; i.e., $(b_{0i}, b_{1i})^T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D})$

Combining:

$$Y_{ij} = \gamma_{0G}(1-G_i) + \gamma_{0B} G_i + \gamma_{1G}(1-G_i) t_{ij} + \gamma_{1B} G_i t_{ij} + b_{0i} + b_{1i} t_{ij} + e_{ij}$$

Can summarize in matrix form... $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$

4. Statistical models for longitudinal data

Linear mixed effects model: $\boldsymbol{Y}_i = \boldsymbol{X}_i \boldsymbol{\gamma} + \boldsymbol{Z}_i \boldsymbol{b}_i + \boldsymbol{e}_i$, $i = 1, \ldots, n$,

$$\boldsymbol{\gamma} = \begin{pmatrix} \gamma_{0G} \\ \gamma_{1G} \\ \gamma_{0B} \\ \gamma_{1B} \end{pmatrix}, \quad \boldsymbol{b}_i = \begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix}, \quad \boldsymbol{Z}_i = \begin{pmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{im_i} \end{pmatrix}$$

$$\boldsymbol{X}_i = \begin{pmatrix} (1-G_i) & (1-G_i)t_{i1} & G_i & G_i t_{ij} \\ \vdots & \vdots & \vdots & \vdots \\ (1-G_i) & (1-G_i)t_{im_i} & G_i & G_i t_{ij} \end{pmatrix}$$

This model is "*subject-specific*" because it acknowledges the *individual subject profiles* through the *random effects* $\boldsymbol{b}_i$

4. Statistical models for longitudinal data

Averaging across the population: If we *average* over subjects (and fluctuations and measurement errors) for each group (value of G_i)

$$\mathrm{E}(\boldsymbol{Y}_i | G_i) = \boldsymbol{X}_i \boldsymbol{\gamma}, \quad \mathrm{var}(\boldsymbol{Y}_i | G_i) = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T + \sigma_f^2 \boldsymbol{H}_i + \sigma_e^2 \boldsymbol{I}_i = \boldsymbol{V}_i$$

so that

$$\boldsymbol{Y}_i | G_i \sim \mathcal{N}(\boldsymbol{X}_i \boldsymbol{\gamma}, \boldsymbol{V}_i)$$

- $\mathrm{E}(\boldsymbol{Y}_i | G_i = 0)$ is the *population average* (*population mean*) for girls
- $\mathrm{E}(\boldsymbol{Y}_i | G_i = 1)$ is the *population average* (*population mean*) for boys
- Which implies

$$\mathrm{E}(Y_{ij} | G_i = 0) = \gamma_{0G} + \gamma_{1G} t_{ij} \quad \text{girls}$$

$$\mathrm{E}(Y_{ij} | G_i = 1) = \gamma_{0B} + \gamma_{1B} t_{ij} \quad \text{boys}$$

4. Statistical models for longitudinal data

Features:

- Questions about "*typical*" *individual behavior* are questions about $\boldsymbol{\gamma}$
- The *covariance matrix* $\boldsymbol{V}_i$ has a particular form with *separate components* for each type of correlation, which the analyst can specify
- Thus, *correlation* is *automatically* taken into account by the model
- *No requirement* for balance

4. Statistical models for longitudinal data

How to specify $\boldsymbol{H}_i$? Common models are borrowed from *time series*

- *Autoregressive structure of order 1*, AR(1), e.g., for dental data

$$\boldsymbol{H}_i = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

- Depends on ρ
- Can be generalized to *unequally-spaced* times

4. Statistical models for longitudinal data

A standard version: Times t_{ij} may be *far apart*

- *Assume* autocorrelation among $e_{f,ij}$ is *negligible* $\Rightarrow$ $\boldsymbol{H}_i = \boldsymbol{I}_i$, an *identity matrix*
- Then $\boldsymbol{V}_i = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T + \underbrace{(\sigma_f^2 + \sigma_e^2)}_{\sigma^2} \boldsymbol{I}_i$
- σ^2 measures variation due to *both* fluctuation and measurement error

Another standard version: *No measurement error* involved

- $\Rightarrow$ $\sigma_e^2 = 0$ so $\boldsymbol{V}_i = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T + \sigma_f^2 \boldsymbol{H}_i = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T + \sigma^2 \boldsymbol{H}_i$
- If *also* times are *far apart* $\boldsymbol{V}_i = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T + \sigma^2 \boldsymbol{I}_i$
- σ^2 measures variation due to *fluctuation*

4. Statistical models for longitudinal data

Back to the dental data: $m_i = 4$

- *Measurement error?* Probably (so $\sigma_e^2 > 0$)
- "*Fluctuations*" in distance? Maybe ($\sigma_f^2 > 0$)
- Measurement error may *dominate* $\Rightarrow$ $\sigma_e^2 >> \sigma_f^2$
- *Autocorrelation*? 2 years is a *long time*
- A reasonable model

$$\boldsymbol{V}_i = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T + \sigma^2 \boldsymbol{I}_i$$

  σ^2 measures primarily measurement error
- Is "*typical*" (mean) slope for girls *different* from that for boys? $\Rightarrow$ Test $\gamma_{1G} = \gamma_{1B}$

4. Statistical models for longitudinal data

Notice: $\boldsymbol{V}_i = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T + \sigma^2 \boldsymbol{I}_i$

- $\boldsymbol{V}_i$ is *not diagonal* in general *even if* autocorrelation is *negligible*!
- $Y_{i1}, \ldots, Y_{im_i}$ are always *correlated* because they all share the *same inherent trajectory* (i.e., b_{0i}, b_{1i})!
- *Correlation* due to *among-subject heterogeneity*
- So *any model* for longitudinal data *must* acknowledge this!

Analysis: *Estimate* parameters (e.g., $\boldsymbol{\gamma}$) and *test hypotheses* (e.g., $\gamma_{0G} = \gamma_{0B}$) by *fitting* this model to the data (coming up...)

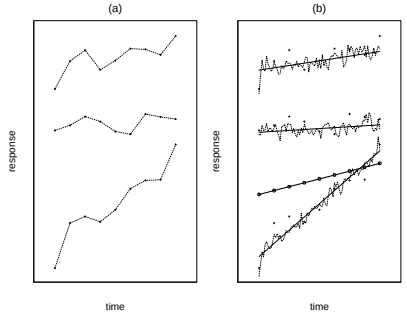
4. Statistical models for longitudinal data

Population-averaged model:

- Model *population behavior* by *modeling* the *population average profiles* $\mathrm{E}(\boldsymbol{Y}_i | G_i)$ *directly*
- Questions are about how *population means* are related over time
- Instead of worrying about separate components of $\mathrm{var}(\boldsymbol{Y}_i | G_i)$ (within- and among-individual sources of correlation), just model their *combined effect* directly
- From the previous slide, this means pick a "*working model*" that will account for *among-subject heterogeneity* at the least!

4. Statistical models for longitudinal data

Conceptualization:



4. Statistical models for longitudinal data

Population-averaged model : For randomly chosen subject i measure Y_{ij} at several times t_{ij} (need not be *the same* for all i)

$$Y_{ij} = \beta_{0G}(1-G_i) + \beta_{0B}G_i + \beta_{1G}(1-G_i)t_{ij} + \beta_{1B}G_i t_{ij} + \epsilon_{ij}$$

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

4. Statistical models for longitudinal data

In matrix form: $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{0G} \\ \beta_{1G} \\ \beta_{0B} \\ \beta_{1B} \end{pmatrix}, \quad \boldsymbol{X}_i = \begin{pmatrix} (1-G_i) & (1-G_i)t_{i1} & G_i & G_i t_{ij} \\ \vdots & \vdots & \vdots & \vdots \\ (1-G_i) & (1-G_i)t_{im_i} & G_i & G_i t_{ij} \end{pmatrix}$$

- $\mathrm{E}(\boldsymbol{Y}_i|G_i) = \boldsymbol{X}_i\boldsymbol{\beta}$ (*population average* for each group)
- Choose a "*working model*" for the covariance matrix $\mathrm{var}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}_i$ that (hopefully) captures the overall combined correlation
- Thus, model is $\mathrm{E}(\boldsymbol{Y}_i|G_i) = \boldsymbol{X}_i\boldsymbol{\beta}$, $\mathrm{var}(\boldsymbol{Y}_i|G_i) = \boldsymbol{\Sigma}_i$
- Can assume *normality* $\boldsymbol{Y}_i|G_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$

4. Statistical models for longitudinal data

How to choose a "working model" $\boldsymbol{\Sigma}_i$?

- Which source of *correlation* dominates?
- *Within-subject autocorrelation* $\Rightarrow$ popular models from *time series*
- *Among-subject heterogeneity* $\Rightarrow$ *compound symmetry* models, e.g., $m_i = 4$

$$\boldsymbol{\Sigma}_i = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$$

4. Statistical models for longitudinal data

Contrasting the models:

Subject-specific: We saw this model implies the *population average* is $\mathrm{E}(\boldsymbol{Y}_i|G_i) = \boldsymbol{X}_i\boldsymbol{\gamma}$ (slide 48)

Population-averaged: We model the *population average directly* as $\mathrm{E}(\boldsymbol{Y}_i|G_i) = \boldsymbol{X}_i\boldsymbol{\beta}$

Result: The models for the *population average* are of the *same form*!

- Thus $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ describe the *same thing*, 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!

Difference: How $\mathrm{var}(\boldsymbol{Y}_i|G_i)$ is represented

4. Statistical models for longitudinal data

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

4. Statistical models for longitudinal data

What about discrete data? Six Cities data

- We observe pairs (Y_{ij}, X_{ij}), $j = 1, \ldots, m_i = 4$ for each child
- $Y_{ij} = 1$ if respiratory infection, $= 0$ if not
  $X_{ij} = 1$ if mother smoking, $= 0$ if not
- $\mathbf{x}_i = (X_{i1}, \ldots, X_{im_i})^T$ is the overall smoking behavior observed
- Natural models (e.g. *logistic*, probit) for *binary response* are *nonlinear*

Subject-specific and population averaged models now have different interpretations. . .

4. Statistical models for longitudinal data

Subject-specific model: Model *individual propensity* for respiratory infection at age j ($Y_{ij} = 1$) when exposed to maternal smoking $\mathbf{x}_i$

- Simplest, popular model

$$\log\left(\frac{P(Y_{ij}=1|\mathbf{x}_i, b_i)}{1-P(Y_{ij}=1|\mathbf{x}_i, b_i)}\right) = \beta_{0i} + \beta_{1i}X_{ij} = \gamma_0 + b_i + \gamma_1 X_{ij}$$

$$\beta_{0i} = \gamma_0 + b_i, \quad \beta_{1i} = \gamma_1, \quad b_i \sim \mathcal{N}(0, D)$$

- $P(Y_{ij} = 1|\mathbf{x}_i, b_i) = \mathsf{E}(Y_{ij}|\mathbf{x}_i, b_i)$ is the probability of infection for *child i in particular* under mother's smoking overall behavior $\mathbf{x}_i$
- Common *assumption* is that this depends *only* on the mother's *current* smoking at j (X_{ij})
- Even in absence of smoking, children are *heterogeneous* in their propensity to have respiratory infections, represented by the *probability distribution* of the b_i

4. Statistical models for longitudinal data

Subject-specific model: Example of a *generalized linear mixed model*

$$\log\left(\frac{P(Y_{ij}=1|\mathbf{x}_i, b_i)}{1-P(Y_{ij}=1|\mathbf{x}_i, b_i)}\right) = \beta_{0i} + \beta_{1i}X_{ij}, = \gamma_0 + b_i + \gamma_1 X_{ij}$$

- β_{0i} is the log odds of respiratory infection *for child i* when mother does not smoke
  $\Rightarrow$ γ_0 is the "*typical*" (mean) value of the log odds for children in the population
- $\beta_{1i} = \gamma_1$ is the change in log odds of respiratory infection when *child i* is exposed to smoking relative to not
- So γ_1 measures change in log odds for *individuals*
- All of $Y_{i1}, \ldots, Y_{im_i}$ depend on b_i $\Rightarrow$ *correlation* due to *among-subject heterogeneity* taken into account

Analysis: Need to *estimate* $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)^T$ in this model

4. Statistical models for longitudinal data

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

$$\log\left(\frac{P(Y_{ij}=1|\mathbf{x}_i)}{1-P(Y_{ij}=1|\mathbf{x}_i)}\right) = \beta_0 + \beta_1 X_{ij}$$

- $P(Y_{ij} = 1|\mathbf{x}_i) = \mathsf{E}(Y_{ij}|\mathbf{x}_i)$ is the probability of respiratory infection at age j *in the population* of children with mother's overall smoking $\mathbf{x}_i$
- Think of this as the *proportion* of children exposed to $\mathbf{x}_i$ who would have a respiratory infection at age j
- Again, that this depends only on X_{ij} is an *assumption*
- This model says *nothing* about individual children (we've *averaged* over them)

4. Statistical models for longitudinal data

Population-averaged model: Thus, with

$$\log\left(\frac{P(Y_{ij}=1|\mathbf{x}_i)}{1-P(Y_{ij}=1|\mathbf{x}_i)}\right) = \beta_0 + \beta_1 X_{ij}$$

- β_0 is the log odds of respiratory infection for the population of children whose mothers don't smoke
- β_1 is the change in log odds of respiratory infection if the population were exposed to smoking relative to not
- Thus, β_0 and β_1 describe what happens "*on average*" in the population (as opposed to what happens for *individuals*)
- *Correlation*? As with continuous data, specify a *working covariance matrix* for var$(\boldsymbol{Y}_i|\mathbf{x}_i)$

Analysis: Need to *estimate* $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ in this model

4. Statistical models for longitudinal data

Population-averaged: $P(Y_{ij} = 1|\mathbf{x}_i) = \mathsf{E}(Y_{ij}|\mathbf{x}_i) = \dfrac{\exp(\beta_0 + \beta_1 X_{ij})}{1 + \exp(\beta_0 + \beta_1 X_{ij})}$

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

Subject-specific: $P(Y_{ij}=1|\mathbf{x}_i, b_i) = \mathsf{E}(Y_{ij}|\mathbf{x}_i, b_i) = \dfrac{\exp(\gamma_0 + b_i + \gamma_1 X_{ij})}{1 + \exp(\gamma_0 + b_i + \gamma_1 X_{ij})}$

- A model *specifically* for the ith child
- We could *average* this over the population by averaging over the b_i to get the *implied population-averaged* model:

$$\int \frac{\exp(\gamma_0 + b_i + \gamma_1 X_{ij})}{1 + \exp(\gamma_0 + b_i + \gamma_1 X_{ij})} \frac{1}{\sqrt{2\pi D}} \exp\left(-\frac{b_i^2}{2D}\right) db_i$$

- This integral (over the $\mathcal{N}(0, D)$ density) is a *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, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ have *different interpretations*

4. Statistical models for longitudinal data

Which perspective/model makes more sense?

- Depends on the *subject matter science* and the *objective*
- A *clinician* deciding between two treatments for a patient would be interested in the difference in response for *an individual*
  ⇒ *subject-specific model*
- For making *public policy* recommendations, what happens *on average* in the population is usually more relevant then what happens to individuals
  ⇒ *population-averaged model*
- If the model is *linear* (as for the *dental data*), can go *either way*! (Either interpretation valid!)
- If an appropriate model is *nonlinear* have to *think carefully*

4. Statistical models for longitudinal data

Sometimes, the choice is clear: Recall the *theophylline PK data*:

- Interested in "*typical values*" and *variation* of k_{ai}, k_{ei}, V_i
- *Nonlinear mixed effects model*

$$Y_{ij} = \frac{k_{ai} D}{V_i(k_{ai}-k_{ei})}\{e^{-k_{ei}t} - e^{-k_{ai}t}\} + e_{ij}, \quad k_{ei} = Cl_i/V_i$$

$$\log k_{ai} = \gamma_1 + b_{k_a,i}, \quad \log Cl_i = \gamma_2 + b_{Cl,i}, \quad \log V_i = \gamma_3 + b_{V,ie}$$

$$\boldsymbol{b}_i = \begin{pmatrix} b_{k_a,i} \\ b_{Cl,i} \\ b_{V,i} \end{pmatrix}, \quad \boldsymbol{b}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D})$$

- Fancier: if i has *weight* w_i, e.g., $\log Cl_i = \gamma_{20} + \gamma_{21} w_i + b_{Cl,i}$ (is *weight important*? $\Leftrightarrow \gamma_{21} = 0$?)
- Clearly this is a *subject-specific* model ⇒ estimate $\boldsymbol{\gamma}$, $\boldsymbol{D}$
- A *population-averaged* model could not address the questions!

5. Implementation

Linear models: *Covariance matrix* plays key role!

Subject-specific models: *Maximum likelihood* estimation of $\boldsymbol{\gamma}$, parameters in $\boldsymbol{V}_i$, based on $\boldsymbol{Y}_i|G_i \sim \mathcal{N}(\boldsymbol{X}_i\boldsymbol{\gamma}, \boldsymbol{V}_i)$

- Given estimates of the parameters in $\boldsymbol{V}_i$, solve

$$\sum_{i=1}^n \boldsymbol{X}^T \widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\gamma}) = \boldsymbol{0} \;\Rightarrow\; \widehat{\boldsymbol{\gamma}} = \left(\sum_{i=1}^n \boldsymbol{X}_i^T \widehat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i\right)^{-1} \sum_{i=1}^n \boldsymbol{X}_i^T \widehat{\boldsymbol{V}}_i^{-1} \boldsymbol{Y}_i$$

- Given $\widehat{\boldsymbol{\gamma}}$, solve another equation for parameters in $\boldsymbol{V}_i$ ("*ML*" or "*REML*")
- Requires iterative numerical algorithm
- SAS `proc mixed`, R `lme()`

5. Implementation

Linear models: *Covariance matrix* plays key role!

Population-averaged models: The same

- Given estimates of the parameters in $\boldsymbol{\Sigma}_i$, solve

$$\sum_{i=1}^n \boldsymbol{X}^T \widehat{\boldsymbol{\Sigma}}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) = \boldsymbol{0} \;\Rightarrow\; \widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \boldsymbol{X}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{X}_i\right)^{-1} \sum_{i=1}^n \boldsymbol{X}_i^T \widehat{\boldsymbol{\Sigma}}_i^{-1} \boldsymbol{Y}_i$$

- Given $\widehat{\boldsymbol{\beta}}$, solve another equation for parameters in $\boldsymbol{\Sigma}_i$
- Not surprising, as *interpretation* is the *same*
- SAS `proc mixed`

5. Implementation

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

Population-averaged models: Solve similar *generalized estimating equations* (*GEEs*) for $\boldsymbol{\beta}$ and parameters in $\boldsymbol{\Sigma}_i$

$$\sum_{i=1}^n \boldsymbol{D}_i^T(\mathbf{x}_i, \boldsymbol{\beta}) \widehat{\boldsymbol{\Sigma}}_i^{-1} \{\boldsymbol{Y}_i - \boldsymbol{\mu}(\mathbf{x}_i, \boldsymbol{\beta})\} = \boldsymbol{0}$$

- E.g., for Six Cities (*binary response*, $\boldsymbol{\mu}(\mathbf{x}_i, \boldsymbol{\beta})$ has jth element

$$\mu_{ij} = \frac{\exp(\beta_0 + \beta_1 X_{ij})}{1 + \exp(\beta_0 + \beta_1 X_{ij})}$$

- $\boldsymbol{\Sigma}_i$ is chosen to have a relevant *correlation* pattern and *diagonal elements* (*variances*) relevant to type of response, e.g., for *binary*

$$\mu_{ij}(1 - \mu_{ij})$$

- SAS `proc genmod`, R `gee( )`

5. Implementation

Subject-specific models: *Maximum likelihood* (*messy*!)

- Maximize in $\boldsymbol{\gamma}$, $\boldsymbol{D}$

$$\prod_{i=1}^n p(\boldsymbol{y}_i|\mathbf{x}_i) = \prod_{i=1}^n \int p(\boldsymbol{y}_i|\mathbf{x}_i, \boldsymbol{b}_i)\, n(\boldsymbol{b}_i; \boldsymbol{0}, \boldsymbol{D})\, d\boldsymbol{b}_i, \quad n(\boldsymbol{b}; \boldsymbol{0}, \boldsymbol{D}) \text{ is } \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}) \text{ density}$$

- $p(\boldsymbol{y}_i|\mathbf{x}_i, \boldsymbol{b}_i)$ is the assumed density of $\boldsymbol{Y}_i$ given $(\mathbf{x}_i, \boldsymbol{b}_i)$
- E.g., for our earlier model for *binary responses* with *no* within-subject autocorrelation

$$p(\boldsymbol{y}_i|\mathbf{x}_i, b_i) = \prod_{j=1}^{m_i} (\mu_{ij}^b)^{y_{ij}} (1 - \mu_{ij}^b)^{1-y_{ij}}, \quad \mu_{ij}^b = \frac{\exp(\gamma_0 + b_i + \gamma_1 X_{ij})}{1 + \exp(\gamma_0 + b_i + \gamma_1 X_{ij})}$$

- *Intractable integration* – integral must be done *numerically* or *approximated* somehow
- SAS `proc nlmixed`, `%glimmix`, `%nlinmix`, R `nlme()`

5. Implementation

In all cases: Standard errors, confidence intervals, hypothesis tests all take into account the assumptions on *correlation*

- If this were ignored, these inferences would be *flawed*!

Why do I need to know all of this? All I want to do is do the analysis!

- In *all cases* the *syntax* of the software is *directly tied* to the *statistical model*!
- Thus, the user *must be clear* about exactly which model s/he wishes to fit

For example...

5. Implementation

SAS `proc mixed`: *Linear mixed effects model* $\boldsymbol{Y}_i = \boldsymbol{X}_i\gamma + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i$

Basic syntax:

```
proc mixed data=dataset method= (ML,REML);
class  classification variables;
model response =  columns of  X / solution;
random columns of  Z / type=  subject=  ;
repeated / type=  subject=  ;
run;
```

- `model` statement specifies rows of $\boldsymbol{X}_i\gamma$
- `random` statement specifies *random effects* and matrix $\boldsymbol{D}$
- `repeated` statement specifies beliefs about e_{ij} (*within-subject* variation) – not needed if autocorrelation is negligible
- `type` options allow choice of matrix, e.g., `un` (unstructured), `ar(1)`, `cs` (compound symmetric), `simple`/`vc` ($\sigma^2\boldsymbol{I}$), ...

5. Implementation

Example: *Dental data* under the assumptions on *fluctuations*, *measurement error* discussed previously

$$Y_{ij} = \gamma_{0G}(1-G_i)+\gamma_{0B}G_i+\gamma_{1G}(1-G_i)t_{ij}+\gamma_{1B}G_it_{ij}+b_{0i}+b_{1i}t_{ij}+e_{ij}$$

$$\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^T + \sigma^2\boldsymbol{I}_i$$

- Because the "within-subject" part of $\boldsymbol{V}_i$ is $\sigma^2\boldsymbol{I}_i$, a `repeated` statement is *not* required, but we show what it would be if we chose to include it

5. Implementation

```
data dental; input child age dist gen @@; oppgen=1-gen;
datalines;
1 8 21 0 1 10 20 0 1 12 21.5 0 1 14 23 0 ...
27 8 22 1 27 10 21.5 1 27 12 23.5 1 27 14 25 1
;

proc mixed method=reml;  * reml is the default;
class child;
model dist = oppgen gen oppgen*age gen*age / noint solution;
random intercept age / type=un subject=child;
repeated / type=simple subject=child;  * could be left out;
run;
```


6. Discussion

Take-away message: Specialized *statistical models* are required for longitudinal data analysis

- Before one can *analyze* longitudinal data, one *must understand* the *models* and their *interpretation*
- Understanding the models is *critical* to understanding on how to use *software*!
- Hence the focus here on *models* rather than *methods*

6. Discussion

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)
- And much more!

6. Discussion

Where to learn more: Some references (there are many others!)

Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*, Springer.

Fitzmaurice, G.M., Laird, N.M., and Ware, J.H. (2004) *Applied Longitudinal Analysis*, Wiley.

Weiss, R.E. (2005) *Modeling Longitudinal Data*, Springer.

Diggle, P.J., Heagerty, P., Liang, K.-Y., and Zeger, S.L. (2002) *Analysis of Longitudinal Data, 2nd Edition*, Oxford University Press.

Molenberghs, G. and Verbeke, G. (2005) *Models for Discrete Longitudinal Data*, Springer.

Davidian, M. and Giltinan, D.M. (1995) *Nonlinear Models for Repeated Measurement Data*, Chapman and Hall/CRC Press.

Vonesh, E.F. and Chinchilli, V.M. (1997) *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, Marcel Dekker.

6. Discussion

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

`http://www.stat.ncsu.edu/~davidian`

(including *lots* of examples of using SAS and R under the ST 732 and ST 762 course web pages!)