Statistical Methods for Correlated Data

A preface to the slides' content

Michele Guindani

Department of Biostatistics
UCLA

A preface – 1

- The slides are a part of a 10-weeks course I taught at UCI last in the Spring 2022
- They provide an introduction to statistical methods for analyzing correlated data. Topics include linear mixed models, non-linear mixed effects models, and generalized estimating equations.
- The content is borrowed mostly from books on the topics.
- The main reference (recommended text for the course) was
- J. Wakefield (2013) *Bayesian and Frequentist Regression Methods*, Chapter 8 and following, Springer.
- Other books heavily employed/borrowed upon are:

1 Fitzmaurice G., Davidian, M., Verbeke, G. and Molenberghs, G. (2009) Handbook of Longitudinal Data Analysis, CRC Press.

A preface – 2

2 Fitzmaurice G., Laird, L., and Ware J. (2004), Applied Longitudinal Data Analysis, Second Edition. Wiley Series in Probability and Statistics.

3 Diggle, P, Heagerty, P, Liang, KY, and Zeger, S. (2002), Analysis of Longitudinal Data, Oxford University Press.

4 McCulloch, C. E., & Neuhaus, J. M. (2001). Generalized linear mixed models. John Wiley & Sons, Ltd.

5 Song, P. (2007) Correlated Data Analysis, Springer.

6 Wood, S.N. (2017) Generalized Additive Models: An Introduction with R, Second Edition, CRC Press.

7 Weiss, R.E. (2005) Modeling Longitudinal Data, Springer texts in Statistics

A preface – 3

- If you notice any other instance where credit is due and was not given, please let me know and I will add it as necessary
- If you are interested in the .tex files and other material (e.g., lab material) for your own course preparation, feel free to reach out directly to me. My contact information can be found here:
- My Website
- My Github

Disclaimer:

Typos and errors are all mine!

Statistical Methods for Correlated Data

Introduction to the course

Michele Guindani

Department of Biostatistics
UCLA

- With linear models and generalized linear models, the models assumed independence (conditional independence) assumptions between observations:

$$Y_1, \ldots, Y_n | X, \beta, \sigma^2 \overset{ind}{\sim} N(X_i\beta, \sigma^2),$$

- where $X_i\beta = \mu_i$ is the regression line, capturing the relationiship between the pairs (x_i, y_i) for all observations $i = 1, \ldots, n$.

- Here, we consider models for dependent data:

1. sampling over time (e.g., measuring individual blood pressure over multiple days)
2. sampling in space (e.g. number of diseased subjects in different counties; *split-plot* design)
3. sampling within clusters (e.g. families)

Longitudinal, time series and clustered data

- *Longitudinal Data Analysis* is concerned with estimating how individual's measurements change over time (e.g. over the duration of a study) and with examining factors that influence heterogeneity among individuals ("what spurs" the changes happening over time)

⇨ goal: characterize change in the response over time as a function of a subset of covariates

⇨ Repeated measure analysis

Longitudinal, time series and clustered data

- Longitudinal data are different than time series data:
  - Time series data are typically high-frequency (high-resoulution) data
  - There are specialized techniques for studying time series data (omit)
- Clustered data: obvservations on the same unit exhibit *residual* dependence due to shared unmeasured variables, after controlling to known regressors.

Why do we need specific analytic tools?

Why do we need specific analytic tools?

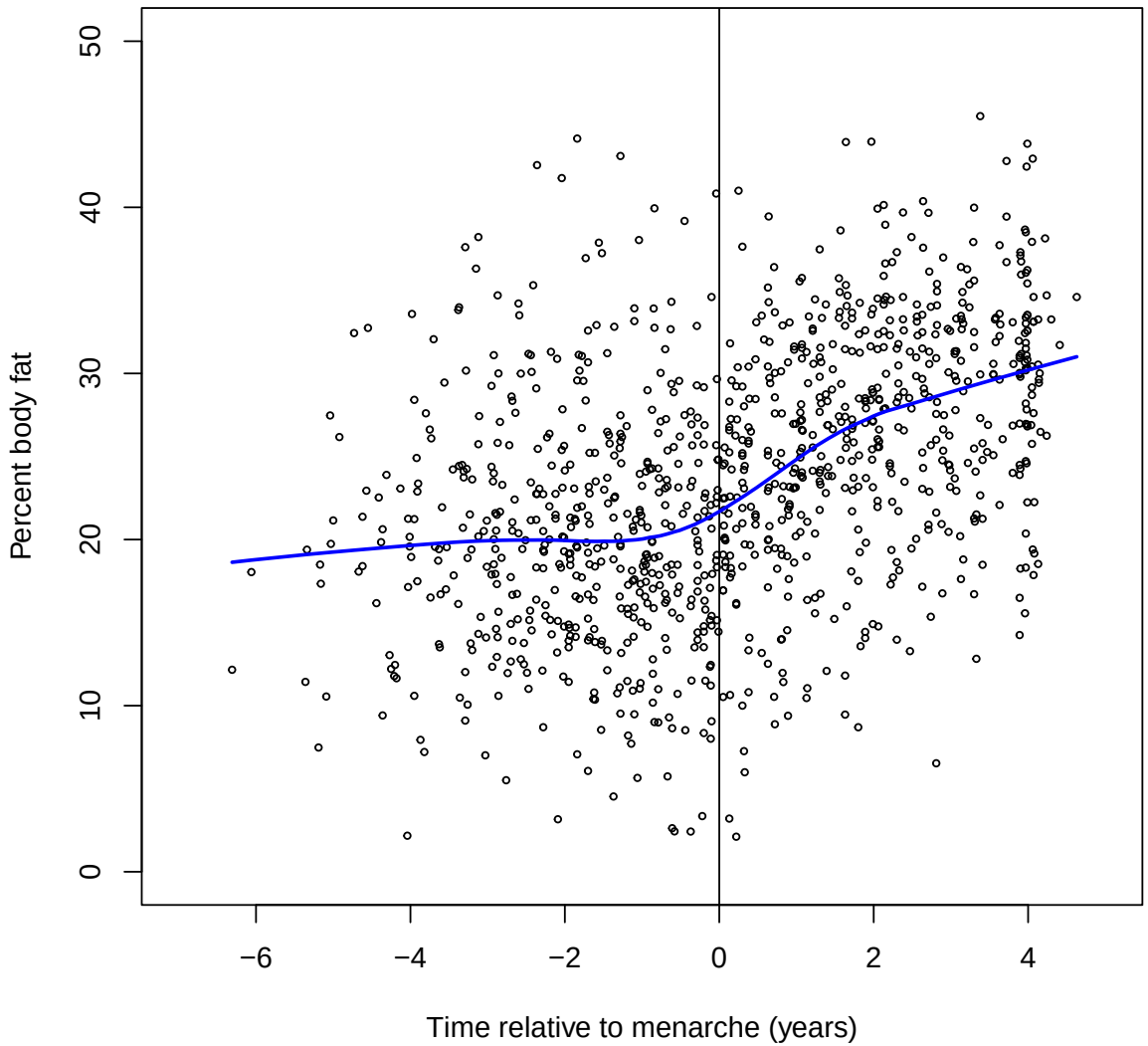
Example (MIT Growth and Development Study)

- Body fatness in girls is thought to increase just before or around menarche (first menstrual period), leveling off approximately 4 years after menarche.
- Researchers are interested in determining the increase in body fatness in girls after menarche.
- How might you design a study to investigate this question?

MIT Growth and Development Study

- Prospective study on body fat accretion in a cohort of 162 girls.
- At the start of the study, all of the girls were pre-menarcheal and non-obese (tricep skinfold thickness less than 85th percentile).
- All girls were followed over time according to a schedule of annual measurements until four years after menarche.
- At each examination, a measure of body fatness was obtained based on bioelectric impedance analysis and a measure of percent body fat was derived.
- A total of 1049 individual percent body fat (PBF) measurements, with an average of 6.4 measurements per subject.

MIT Growth and Development Study



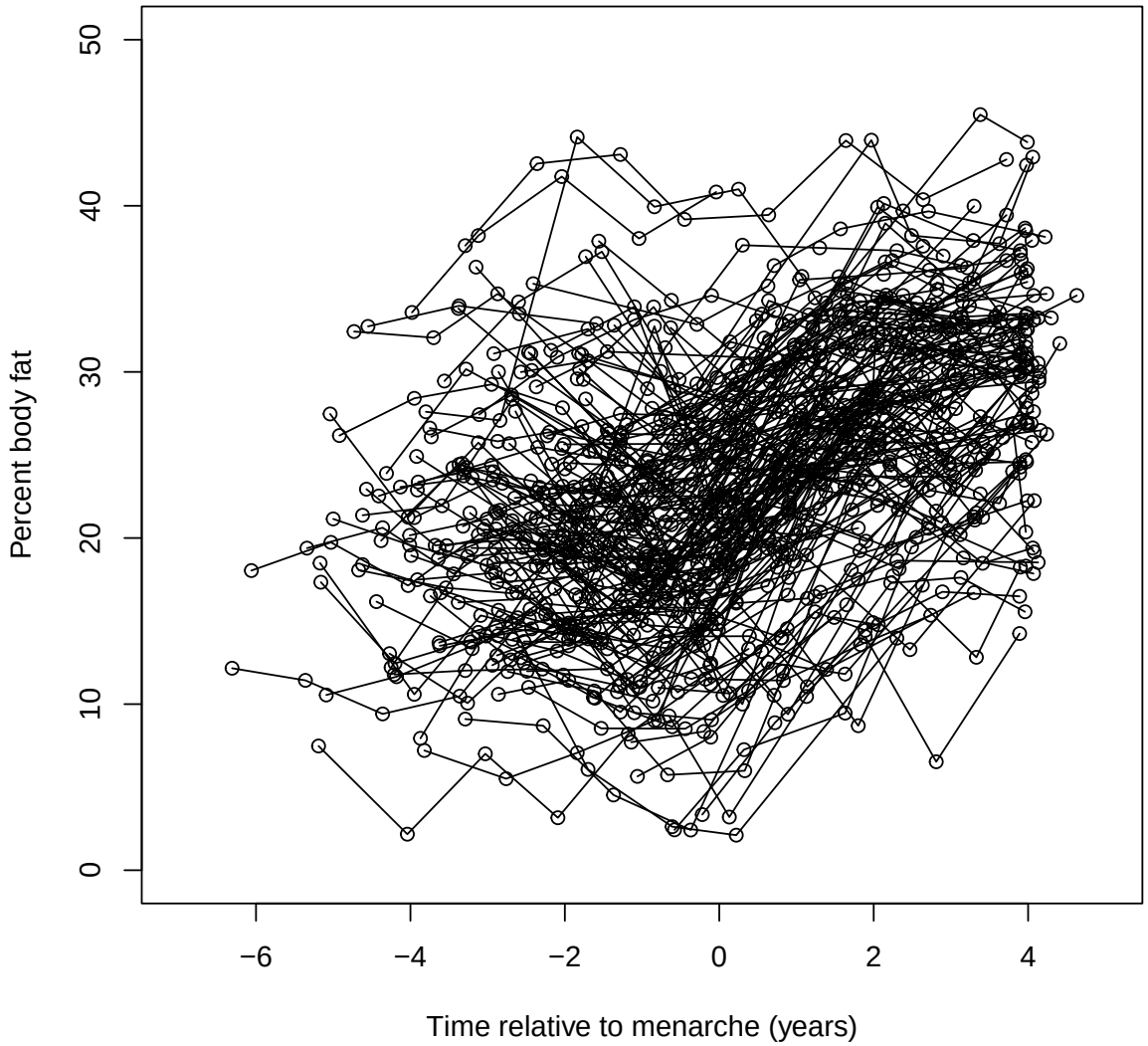
Naive solution

```
summary(lm(formula=PBF~Time.M))

##
## Call:
## lm(formula = PBF ~ Time.M)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.5499  -4.5766   0.2428   4.7401  23.5149
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.33576    0.22157  105.32   <2e-16 ***
## Time.M       1.47321    0.09459   15.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.153 on 1047 degrees of freedom
## Multiple R-squared:  0.1881,Adjusted R-squared:  0.1873
## F-statistic: 242.6 on 1 and 1047 DF,  p-value: < 2.2e-16
```

- What is wrong with this analysis?

Time plot with joined line segments (line plot):



A better comparison

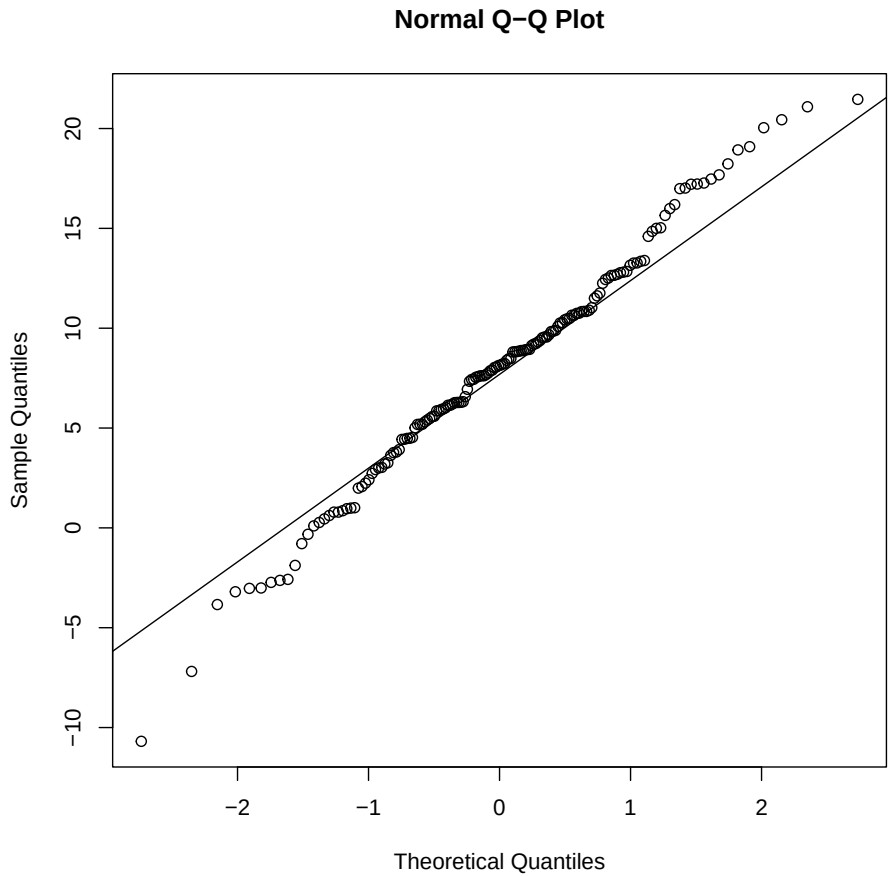
- Paired t-test - Look at each girl's first (pre-menarche) and last (post-menarche) PBF measurements and calculate the difference in PBF for each girl. Is there a significant positive mean gain in PBF?

```
# Paired t-test:
t.test(PostPBF,PrePBF,paired=TRUE,alternative="greater")

##
##  Paired t-test
##
## data:  PostPBF and PrePBF
## t = 17.521, df = 159, p-value < 2.2e-16
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  7.195989      Inf
## sample estimates:
## mean difference
##        7.946375
```

Assumptions of the t-test

Normality of the differences? Data show heavy-tailed distributions



But robust for large samples - we have N = 162

- Independence? Ok since each difference was measured on a different girl.

Limitations of the t-test

- Not all girls' measurements were taken at the same time relative to menarche:
- Premenarche measurements ranged from 0.13 to 6.31 years prior to menarche
- Post-menarche measurements ranged from 0.38 to 4.63 years after menarche.
- Two girls had their last measurements taken at 0.03 and 0.04 years prior to menarche - had to throw out those two data points.
- Only tests if there was a significant mean difference between first and last measurements; can't model how PBF changes over time.

⇨ We would like to be able to incorporate the positive correlation between repeated measurements on the same individual into our model

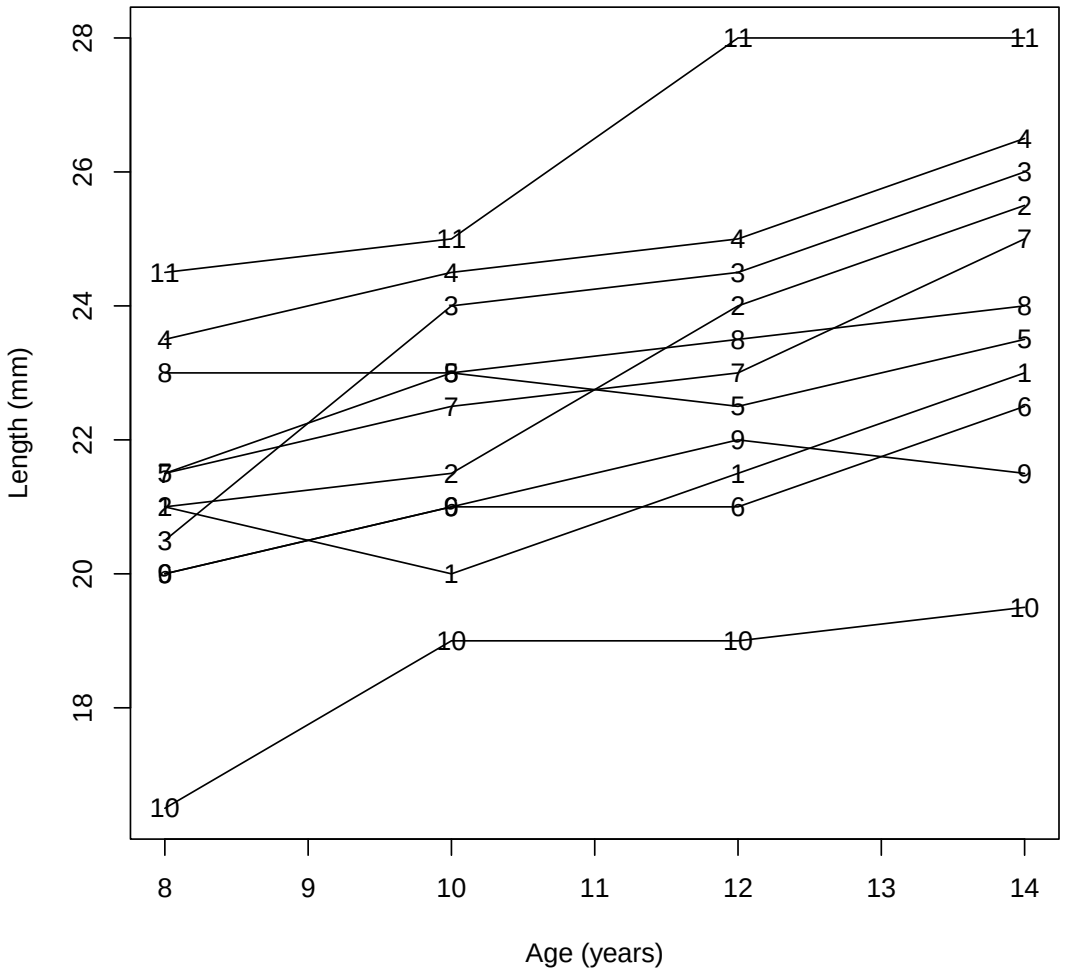
A second example: Dental Growth Curves

We consider dental measurements of the distance in millimeters from the center of the pituitary gland to the pteryo-maxillary fissure are for 11 girls and 16 boys recorded at the ages of $8, 10, 12,$ and 14 years. For now, we concentrate on the data from the girls only.

```
library(nlme)
data(Orthodont)
head(Orthodont, n=15)
```

```
## Grouped Data: distance ~ age | Subject
##    distance age Subject  Sex
## 1      26.0   8     M01 Male
## 2      25.0  10     M01 Male
## 3      29.0  12     M01 Male
## 4      31.0  14     M01 Male
## 5      21.5   8     M02 Male
## 6      22.5  10     M02 Male
## 7      23.0  12     M02 Male
## 8      26.5  14     M02 Male
## 9      23.0   8     M03 Male
## 10     22.5  10     M03 Male
## 11     24.0  12     M03 Male
## 12     27.5  14     M03 Male
## 13     25.5   8     M04 Male
## 14     27.5  10     M04 Male
## 15     26.5  12     M04 Male
```

Figure 8.1 in the textbook shows the so-calles "spaghetti-plots", i.e. the profiles of distance measurements collected on the 11 girls over the years.



The slopes look quite similar, though there is clearly between-girl variability in the intercepts.

- Potential objectives of the analysis of these data:
  1. Population inference, in which we describe the average growth as a function of age, for the population from which the sample of children were selected.
  2. Assessment of the within- to between-child variability in growth measurements.
  3. Individual-level inference, either for a child in the sample, or for a new unobserved child (from the same population) ("growth chart")

Linear Mixed models and Marginal models

- We will discuss mixed effects models which contain both fixed effects that are shared by all individuals and random effects that are unique to particular individuals and are assumed to arise from a distribution.

  The linear mixed effects model allows the estimation of a single curve for each girl

  By marginalizing over the random effects, we can obtain marginal or population-wide inference: let Y_{ij} denote the j th measurement taken at time t_j on the i th child, $i = 1, \ldots, m = 11$,
  $j = 1, \ldots, n_i = 4$.
  Then, consider the model:

$$\mathrm{E}\left[Y_{ij}\right] = \beta_0^{\mathrm{M}} + \beta_1^{\mathrm{M}} t_j$$

  with β_0^{M} and β_1^{M} *marginal* intercept and slope parameters.

Linear Mixed models and Marginal models

The residuals,

$$e_{ij}^{\mathrm{M}} = Y_{ij} - \beta_0^{\mathrm{M}} - \beta_1^{\mathrm{M}} t_j$$

$i = 1, \ldots, 11; j = 1, \ldots, 4$, denote marginal residuals.

Due to the dependence of observations on the same girl, we would not expect the marginal residuals to be independent.

Let

$$\begin{bmatrix} \sigma_1 & & & \\ \rho_{12} & \sigma_2 & & \\ \rho_{13} & \rho_{23} & \sigma_3 & \\ \rho_{14} & \rho_{24} & \rho_{34} & \sigma_4 \end{bmatrix}$$

represent the standard deviation/correlation matrix of the residuals.

Linear Mixed models and Marginal models

In the matrix before,

$$\sigma_j = \sqrt{\text{var}\left(e_{ij}^{\text{M}}\right)}$$

denotes the standard deviation of the dental length at time t_j and

$$\rho_{jk} = \frac{\text{cov}\left(e_{ij}^{\text{M}}, e_{ik}^{\text{M}}\right)}{\sqrt{\text{var}\left(e_{ij}^{\text{M}}\right)\text{var}\left(e_{ik}^{\text{M}}\right)}}$$

denotes the correlation between residual measurements taken at times t_j and t_k on the same girl, $j \neq k, j, k = 1, \ldots, 4$. We assume that these standard deviations and correlations are constant across all girls.

We fit the marginal model to these data and then empirically estimates the entries of the correlation matrix as

$$\begin{bmatrix} 2.12 \\ 0.831.90 \\ 0.860.902.36 \\ 0.840.880.952.44 \end{bmatrix} \tag{1}$$

showing a clear correlation between residuals at different ages on the same girl.

⇨ Hence, using methods for independent data that assume that within-girl correlations are zero will clearly give inappropriate standard errors/uncertainty estimates for $\widehat{\beta}_0^{\mathrm{M}}$ and $\widehat{\beta}_1^{\mathrm{M}}$.

Take-away points

Fitting marginal models allows population-wide inferences: it allows the direct assessment of the average responses at different times.

Marginal models require minimal assumptions. We only used information about the mean function (and correlations) in the previous model

Marginal models require obtaining a realiable estimation of the correlation functions, in order to model within-individual correlations

Linear Mixed models and Marginal models

As a counterpoint to marginal models, linear mixed models allow to estimate a curve for each child while "borrowing strength" in the estimation across children

Note: One could consider a separate estimate for each curve, where each child's profile is estimated separately (all fixed effects, no random part).

The linear mixed model approach allows to estimate population-level parameters (fixed effects) while accommodating individual variation.

In the frequentist setting, random effects are typically seen as a convenient tool, to model within- and between- individual variability. Often no interest in the "individual" random effects' value.
In the Bayesian framework, random effects are latent variables, with a prior, also used to model unobserved correlation. However, there is often a real interest in obtaining posterior estimates of these latent variables.

What are correlated data?

- We will be looking at "clustered" data - Observations within each "cluster" are correlated with each other.

  Positive correlation ⇨ large measurements tend to cluster with large measurements.
  Negative correlation ⇨ large measurements tend to cluster with small measurements.

- Examples of "clusters"?

  Longitudinal data: Repeated measurements taken on the same individual over time.
  Measurements taken on both a mother and daughter.
  Measurements taken on all individuals in a household.

Cross-sectional vs Longitudinal

- In a cross-sectional study, measurements are obtained at only a single point in time.
  - ⇨ It is not possible to assess individual changes across time.

- In a longitudinal study, participants are measured repeatedly throughout the duration of the study
  - ⇨ Permits direct assessment of changes in the response variable over time.
  - Participants or units being studied = *individuals or subjects*. Individuals are measured repeatedly at different times or occasions. Times need not be equally spaced.
  - ⇨ Thus the defining feature of a longitudinal study is that two or more observations of the response variable, taken at different times, are made on at least some of the study participants.

Terminology of Longitudinal studies

- If all individuals have the same number of repeated measurements obtained at a common set of occasions, we say the study is "balanced" over time.

- If repeated measurements are not obtained at a common set of occasions (or individuals have differing numbers of measurements), the study is "unbalanced" over time.

  Common when study is retrospective (e.g., data obtained from medical databases) or when times defined relative to some individual benchmark event, e.g., menarche study.

- If there are missing data (an intended measurement could not be obtained), the data set is called "incomplete."

Missing data are the rule, not the exception, in longitudinal studies in the health sciences. For example, study participants do not always appear for a scheduled observation, or they may simply leave the study before its completion. When some observations are missing, the data are necessarily unbalanced over time, since not all individuals have the same number of repeated measurements obtained at a common set of occasions.

Goals of Longitudinal Studies

- There are two goals in longitudinal data analysis:
- Assess within-individual (intra-individual) changes in the response variable.

  How do we characterize the change in the response variable over time?
- Assess between-individual (inter-individual) changes in the response variable.

  Are the "response trajectories" of individuals related to certain covariates?
- Cross-sectional studies are only able to assess between-individual variation.

Correlation in Longitudinal Data

Nature of correlations among repeated measures taken on one individual:

1. positive
2. decrease with increasing time separation
3. rarely approach zero for pairs of measurements taken far apart in time
4. rarely approach one for pairs of measurements taken very closely together in time

Sources of Variation: (1) Between-Subject variability

- Between-subject heterogeneity in mean response:
  - Some individuals consistently respond higher than average, and others lower.
    e.g., annual income, daily caloric intake, systolic blood pressure
  - Induces a positive correlation between repeated measurements

- Between-subject heterogeneity in response trajectory:
  - Some individuals improve more quickly than others, and some may worsen.
    e.g., CD4 lymphocyte counts after antiviral treatment in AIDS patients, or rate of increase in annual income
  - Often induces decreasing correlations with increasing time separation, e.g., scores at times 1 and 4 often less correlated than scores at times 1 and 2.

- In statistical models, between-individual variability can be accounted for by the introduction of individual-specific random effects (e.g., randomly varying intercepts and slopes)

(2) Within-Subject variability and (3) Measurement Error

- Within-subject biological variation:
  - Repeated measures are realizations of some biological process operating within the individual.

    e.g., weight, systolic blood pressure, serum cholesterol
  - Serial correlation: a stronger correlation for measurements that are closer together in time:

    underlying biological process (or combination of processes) that changes through time in a relatively smooth and continuous fashion.
- Measurement error
  - Not to be confused with within-subject biological variation.
  - May shrink the correlation among repeated measures closer to zero.

Sources of Variation in Longitudinal Data

Graphical representation of the three sources of variability in longitudinal data for two hypothetical individuals:

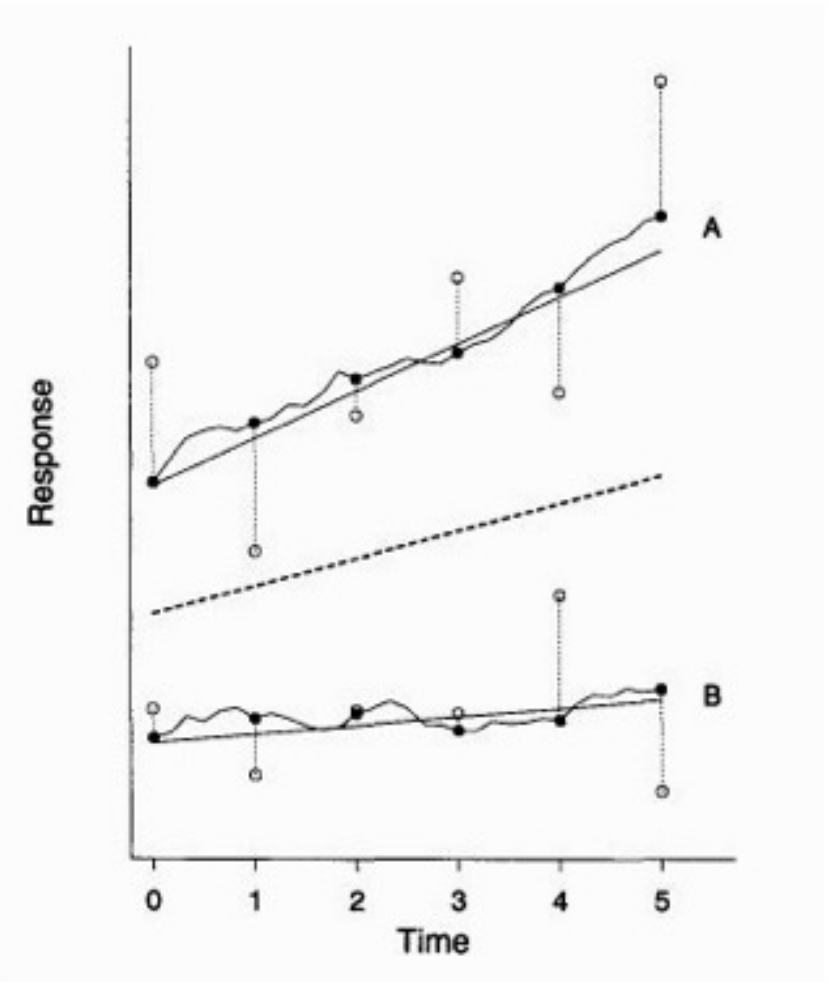
1. Between-individual heterogeneity
2. Within-individual biological variation
3. Measurement error

● denotes repeated measure free of measurement error,
○ denotes observed repeated measure with measurement error.

Solid line represents true individual response trajectory (free of biological variation);

jagged curve is within individual biological variation from solid line.

Dotted line is average true response trajectory between the two respondents



What if we discarded the nature of the data?

- Consider a response variable that both changes over time and varies among subjects. Examples include age, blood pressure, or weight. We start with a simple model:

$$Y_{ij} = \beta_0 + \beta x_{ij} + \epsilon_{ij}, \quad j = 1, \ldots, n; \quad i = 1, \ldots, m \tag{2}$$

for measurements $Y_{i,j}$ collected on n occasions (index j) over m individuals (index i).

What if we discarded the nature of the data?

- Consider a response variable that both changes over time and varies among subjects. Examples include age, blood pressure, or weight. We start with a simple model:

$$Y_{ij} = \beta_0 + \beta x_{ij} + \epsilon_{ij}, \quad j = 1, \ldots, n; \quad i = 1, \ldots, m \tag{2}$$

for measurements $Y_{i,j}$ collected on n occasions (index j) over m individuals (index i).

- We can re-express the previous model as follows:

$$Y_{ij} = \beta_0 + \beta x_{i1} + \beta \left(x_{ij} - x_{i1} \right) + \epsilon_{ij}$$

which makes explicit the assumption: the cross-sectional effect due to x_{i1} is the same as the longitudinal effect represented by $x_{ij} - x_{i1}$ on the right-hand side. This assumption is rather a strong one and doomed to fail in many studies.

- The model can be modified by allowing each person to have their own intercept, β_{0i}, i.e. by replacing $\beta_0 + \beta x_{i1}$ with 1 with β_{0i}:

$$Y_{ij} = \beta_{0i} + \beta \left(x_{ij} - x_{i1} \right) + \epsilon_{ij}$$

- This is also an extreme case, since we allow the baseline to be different for each person.

An alternative modeling

- An intermediate (and more useful) case is to assume a model of the form:

$$Y_{ij} = \beta_0 + \beta_{\mathrm{C}} x_{i1} + \beta_{\mathrm{L}} \left(x_{ij} - x_{i1} \right) + \epsilon_{ij} \tag{3}$$

- The inclusion of $\beta_C \, x_{i1}$ allows both cross-sectional and longitudinal effects to be examined separately.

- We can also use this form to test whether the cross-sectional and longitudinal effects of particular explanatory variables are the same, that is, whether $\beta_{\mathrm{C}} = \beta_{\mathrm{L}}$.

- x_{i1} can be seen as a confounding variable whose absence may bias our estimate of the true longitudinal effect.

Bias

- If we use model (2) the least-squares estimate of β is

$$\hat{\beta} = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x})(y_{ij}-\bar{y}) \,/\, \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2$$

  where $\bar{x} = \sum_{ij} x_{ij}/(nm)$ (average of covariate measurements, e.g. treatment, across all indiv. and times) and $\bar{y} = \sum_{ij} y_{ij}/(nm)$.

- However if the true model is (3), then

$$\mathrm{E}(\hat{\beta}) = \beta_{\mathrm{L}} + \frac{\sum_{i=1}^{m} n\,(x_{i1}-\bar{x}_1)(\bar{x}_i-\bar{x})}{\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2}\,(\beta_{\mathrm{C}}-\beta_{\mathrm{L}})$$

  where $\bar{x}_i = \sum_j x_{ij}/n$ (average across occasions for each individual) and $\bar{x}_1 = \sum_i x_{i1}/m$ (average of all individual covariates at time 1).
- Hence, the cross-sectional estimate $\hat{\beta}$, which assumes $\beta_{\mathrm{L}} = \beta_{\mathrm{C}}$, is a biased estimate of β_{L} and is unbiased only if $\beta_{\mathrm{L}} = \beta_{\mathrm{C}}$ or the variables $\{x_{i1}\}$ and $\{\bar{x}_i\}$ are orthogonal to each other.
- The direction of the bias in $\hat{\beta}$ as an estimate for the longitudinal effect, β_{L}, depends upon the correlation between x_{i1} and $\bar{x}_i$.

Statistical Methods for Correlated Data

The Efficiency of Longitudinal designs

Michele Guindani

Department of Biostatistics
UCLA

The Efficiency of Longitudinal designs

The Efficiency of Longitudinal designs

- Designs that collect dependent data can be very efficient: for example, in a longitudinal data setting, applying different treatments to the same patient over time can be very beneficial, since each patient acts as his/her own control.

- Suppose we want to compare two treatments, coded as -1 and $+1$, with four measurements total.

- In a cross-sectional study, a single measurement is taken on each of four individuals ($n = 4$) with two ($i = 1, 2$) assigned to treatment (-1) and two assigned ($i = 3, 4$) to treatment ($+1$). We can consider the regression model

$$Y_{i1} = \beta_0 + \beta_1 x_{i1} + \epsilon_{i1}$$

with $i = 1, \ldots, m = 4$ and x_{i1} indicating the treatment

- The treatment effect is then

$$\mathrm{E}\left[Y_1|x=1\right]-\mathrm{E}\left[Y_1|x=-1\right]=2\beta_1$$

and the (unbiased) ordinary least squares (OLS) estimators are

$$\widehat{\beta}_0^{\mathrm{c}}=\frac{\sum_{i=1}^{4}Y_{i1}}{4},\quad \widehat{\beta}_1^{c}=\frac{Y_{31}+Y_{41}-(Y_{11}+Y_{21})}{4}$$

and the variance of the treatment estimator is

$$\mathrm{var}\left(\widehat{\beta}_1^{\mathrm{c}}\right)=\frac{\sigma^2}{4}$$

- For the longitudinal study, we assume to have two observations on each of two individuals (for a total , again, of 4 measurements):

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \delta_{ij}$$

where b_i (individual-specific effect) and δ_{ij} (measurement error) are independent and

$$\mathrm{E}\left[\delta_{ij}\right] = 0, \mathrm{var}\left(\delta_{ij}\right) = \sigma_\delta^2, \mathrm{E}\left[b_i\right] = 0, \mathrm{var}\left(b_i\right) = \sigma_0^2$$

- Then, marginalizing with respect to b_i,

$$\begin{aligned}
E\left(Y_{ij}\right) &= \beta_0 + \beta_1 x_{ij} \\
V\left(y_{ij}\right) &= \mathrm{Var}\left(b_i + \delta_{ij}\right) = \sigma_0^2 + \sigma_\delta^2 = \sigma^2 \\
Cov\left(y_{ij}, y_{ik}\right) &= Cov\left(b_i + \delta_{ij}, b_i + \delta_{ik}\right) \\
&= Var\left(b_i\right) + Cov\left(\delta_{ij}, \delta_{ik}\right) + Cov\left(\delta_{ij}, b_i\right) + Cov\left(b_i, \delta_{ik}\right) \\
&= \sigma_0^2
\end{aligned}$$

- Using a vector notation, $\boldsymbol{Y} = [Y_{11}, Y_{12}, Y_{21}, Y_{22}]$, and $\text{var}(\boldsymbol{Y}) = \sigma^2 \boldsymbol{R}$, with

$$\boldsymbol{R} = \begin{bmatrix} 1 & \rho & 0 & 0 \\ \rho & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{bmatrix}$$

where $\rho = \sigma_0^2/\sigma^2$ is the correlation between observations on the same individual.

- Using the marginal model, we can use generalized least squares to obtain the unbiased estimator

$$\widehat{\boldsymbol{\beta}}^{\text{L}} = \left(\boldsymbol{x}^{\text{T}} \boldsymbol{R}^{-1} \boldsymbol{x}\right)^{-1} \boldsymbol{x}^{\text{T}} \boldsymbol{R}^{-1} \boldsymbol{Y}$$

with

$$\text{var}\left(\widehat{\boldsymbol{\beta}}^{\text{L}}\right) = \left(\boldsymbol{x}^{\text{T}} \boldsymbol{R}^{-1} \boldsymbol{x}\right)^{-1} \sigma^2$$

- It is easy to show that the efficiency of the longitudinal design is

$$\frac{\text{var}\left(\widehat{\beta}_1^{\text{L}}\right)}{\text{var}\left(\widehat{\beta}_1^{\text{c}}\right)} = \frac{\left(1-\rho^2\right)}{1-\rho\left(x_{11}x_{12}+x_{21}x_{22}\right)/2}$$

- Two cases: Assuming constant treatment for each individual for the two measurements : $x_{11} = x_{12} = -1, x_{21} = x_{22} = 1$, then

$$\frac{\text{var}\left(\widehat{\beta}_1^{\text{L}}\right)}{\text{var}\left(\widehat{\beta}_1^{\text{c}}\right)} = 1+\rho$$

⇨ the cross-sectional study is preferable in the usual situation in which observations on the same individual display positive correlation (benefit from adding more subjects)

- It is easy to show that the efficiency of the longitudinal design is

$$\frac{\text{var}\left(\widehat{\beta}_1^{\text{L}}\right)}{\text{var}\left(\widehat{\beta}_1^{\text{c}}\right)} = \frac{\left(1-\rho^2\right)}{1-\rho\left(x_{11}x_{12}+x_{21}x_{22}\right)/2}$$

- Two cases: Assuming varying treatment for each individual for the two measurements: $x_{11} = x_{22} = 1, x_{12} = x_{21} = -1$ then

$$\frac{\text{var}\left(\widehat{\beta}_1^{\text{L}}\right)}{\text{var}\left(\widehat{\beta}_1^{\text{c}}\right)} = 1-\rho$$

⇨ the longitudinal study is more efficient when $\rho > 0$, because each individual is acting as his/her own control.

⇨ These results extend to the case of time-varying covariates.

Statistical Methods for Correlated Data

Linear Mixed Models

Michele Guindani

Department of Biostatistics
UCLA

Linear Mixed Models

- Basic idea: The measurements on each unit (individual) can be explained by a regression model characterized by a combination of

Linear Mixed Models

- Basic idea: The measurements on each unit (individual) can be explained by a regression model characterized by a combination of
  - Fixed effects: common to all units in the population
  - Random effects: unit-specific perturbations

Linear Mixed Models

- Basic idea: The measurements on each unit (individual) can be explained by a regression model characterized by a combination of
  - Fixed effects: common to all units in the population
  - Random effects: unit-specific perturbations
- Let $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{in_i}]^{\mathrm{T}}, i = 1, \ldots, m$ denote the measurements over n_i occasions on subject $i = 1, \ldots, m$.
- Let $\boldsymbol{x}_{ij} = [1, x_{ij1}, \ldots, x_{ijk}]$ be a $(k+1) \times 1$ be vector of covariates measured at each occation j, $\boldsymbol{x}_i = [\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}]$ be the design matrix for unit (individual) i
- Let $z_{ij} = [1, z_{ij1}, \ldots, z_{ijq}]^{\mathrm{T}}$ be a $(q+1) \times 1$ vector of variables (e.g., a subset of $\boldsymbol{x}_{ij}$, $\boldsymbol{z}_i = [\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{in_i}]^{\mathrm{T}}$ be the design matrix for unit i.

Linear Mixed Models

- A linear mixed model can be described by two levels (or stages):

  1. Conditional model for the response:

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

  where ϵ_i is an $n_i \times 1$ zero-mean vector of error terms, $i = 1, \ldots, m$

Linear Mixed Models

- A linear mixed model can be described by two levels (or stages):

1. Conditional model for the response:

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

where ϵ_i is an $n_i \times 1$ zero-mean vector of error terms, $i = 1, \ldots, m$

2. Assumptions on the random components:

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{\epsilon}_i\right] = \mathbf{0}, \mathrm{var}\left(\boldsymbol{\epsilon}_i\right) &= \boldsymbol{E}_i(\boldsymbol{\alpha}) \\ \mathrm{E}\left[\boldsymbol{b}_i\right] = \mathbf{0}, \mathrm{var}\left(\boldsymbol{b}_i\right) &= \boldsymbol{D}(\boldsymbol{\alpha}) \\ \mathrm{cov}\left(\boldsymbol{b}_i, \boldsymbol{\epsilon}_{i'}\right) = \mathbf{0}, \quad & i, i' = 1, \ldots, m \end{aligned}$$

where $\boldsymbol{\alpha}$ is an $r \times 1$ vector containing the collection of variance-covariance parameters. Further, $\mathrm{cov}\left(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_{i'}\right) = 0$ and $\mathrm{cov}\left(\boldsymbol{b}_i, \boldsymbol{b}_{i'}\right) = 0$, for $i \neq i'$

Linear mixed models

- Averaging over the random effects, we obtain

$$\begin{aligned} \mathrm{E}\left[\boldsymbol{Y}_i\right] &= \boldsymbol{x}_i\boldsymbol{\beta} \\ \mathrm{var}\left(\boldsymbol{Y}_i\right) &= \boldsymbol{V}_i(\boldsymbol{\alpha}) \\ &= \boldsymbol{z}_i\boldsymbol{D}(\boldsymbol{\alpha})\boldsymbol{z}_i^{\mathrm{T}} + \boldsymbol{E}_i(\boldsymbol{\alpha}) \end{aligned}$$

for $i = 1, \ldots, m$, so that $\boldsymbol{V}_i(\boldsymbol{\alpha})$ is an $n_i \times n_i$ matrix.

- The random effects have induced a structure of dependence across within-individual measurements. Because the random effects $\mathbf{b}_i$ are fixed by group, not varying with observation, the within-group observations share the same random effects and are, therefore, correlated. Note, also, that the diagonal elements of $\mathbf{Z}_i\boldsymbol{D}(\boldsymbol{\alpha})\mathbf{Z}_i'$ need not be constant, so that the previous set-up can also accommodate heteroscedasticity.

- Still,

$$\mathrm{cov}\left(\boldsymbol{Y}_i, \boldsymbol{Y}_{i'}\right) = \mathbf{0}$$

Linear Mixed Models

- The within-group error contribution to the marginal covariance matrix is given directly by $\boldsymbol{E}_i(\boldsymbol{\alpha})$, which can be non-diagonal (correlation) and have different diagonal elements (heteroscedasticity).

- The $\boldsymbol{E}_i(\boldsymbol{\alpha})$ matrices are typically assumed to depend on i only through their dimensions, being parameterized by a fixed, generally small, set of parameters (e.g., $\sigma^2 I_n$, in case of independent errors, or through an AR(1) structure (Box et al., 1994).

- Even though the random effects are useful and intuitive quantities to represent between- group differences in the coefficients, they are not observable in practice. Therefore, likelihood estimation and inference in the frequentist setting generally rely on the marginal distribution of the observed response vectors y_i.

Example 1: ANOVA model

- *One-way random effects model* A model is called a (pure) random effects model if the only fixed effect is an unknown mean, i.e.

$$y_{ij} = \mu + b_i + \epsilon_{ij}$$

where $b_i \overset{i.i.d.}{\sim} N(0, \sigma_0^2)$ and $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2)$.

- Of course, $\boldsymbol{b} \sim N(0, \sigma_0^2 \, I_n)$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma_\epsilon^2 \, I_n)$. In this case, $\boldsymbol{\alpha} = (\sigma_0^2, \sigma_\epsilon^2)$.

Example 1: ANOVA model

- *One-way random effects model* A model is called a (pure) random effects model if the only fixed effect is an unknown mean, i.e.

$$y_{ij} = \mu + b_i + \epsilon_{ij}$$

where $b_i \overset{i.i.d.}{\sim} N(0, \sigma_0^2)$ and $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2)$.

- Of course, $\boldsymbol{b} \sim N(0, \sigma_0^2 I_n)$ and $\boldsymbol{\varepsilon} \sim N(0, \sigma_\epsilon^2 I_n)$. In this case, $\boldsymbol{\alpha} = (\sigma_0^2, \sigma_\epsilon^2)$.

- *Two-way random effects model*:

$$y_{ij} = \mu + \xi_i + \eta_j + \epsilon_{ij}$$

with $\xi_i \sim N\left(0, \sigma_1^2\right), \eta_j \sim N\left(0, \sigma_2^2\right)$; and ϵ_{ij}'s are independent errors distributed as $N\left(0, \tau^2\right)$. Again, assume that the random effects and errors are independent.

Random intercept model

Random intercept model

- This is another very simple and much used model in longitudinal data analysis, the main difference being that the observations are allowed to depend on covariates:

$$y_{ij} = x'_{ij}\beta + b_i + \varepsilon_{ij}$$

- b_i describes how the mean response of individual i deviates for the population average, $b_i \sim N\left(0, \sigma_b^2\right)$.
- ε_{ij} is the measurement or sampling error, $\varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2\right)$.
- This is the model behind the so-called "Repeated measures ANOVA":

$$\text{Var}\left(y_{ij}\right) = \text{Var}\left(x'_{ij}\beta + b_i + \varepsilon_{ij}\right) = \sigma_{b_i}^2 + \sigma_\varepsilon^2$$
$$\text{Cov}\left(y_{ij}, y_{ik}\right) = Cov\left(x'_{ij}\beta + b_i + \varepsilon_{ij}, x'_{ik}\beta + b_i + \varepsilon_{ik}\right) = \sigma_b^2$$

- We will call this covariance structure a compound symmetry covariance structure (soon).

Random intercept and Slope model

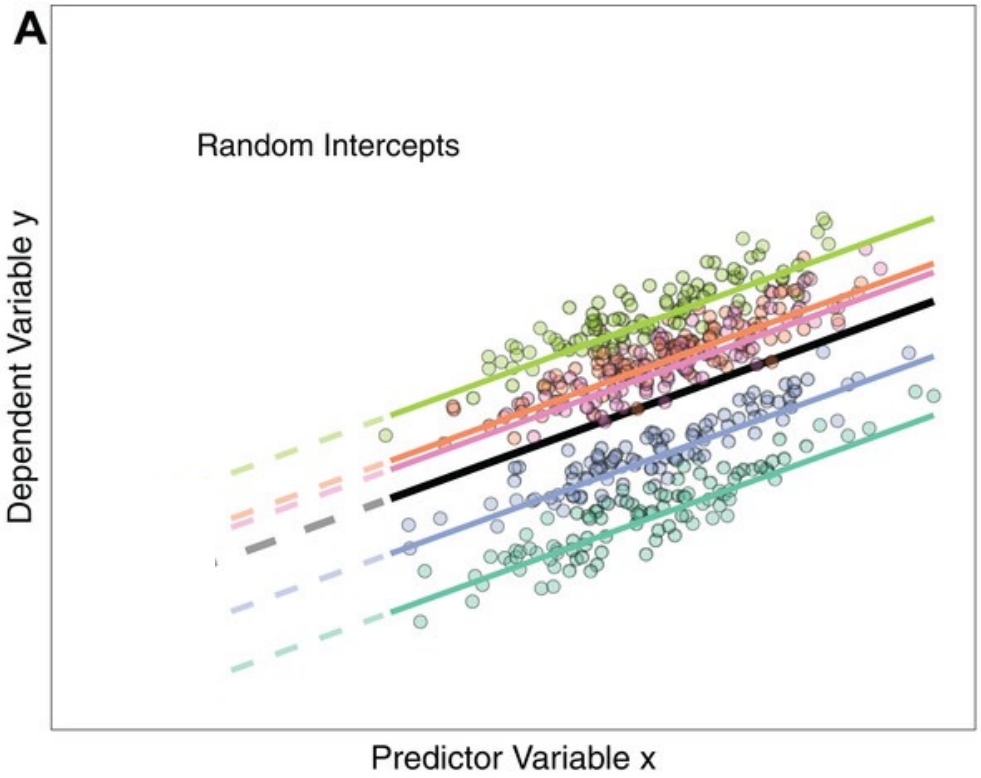
- The mean response changes only as a function of time (no other covariates)

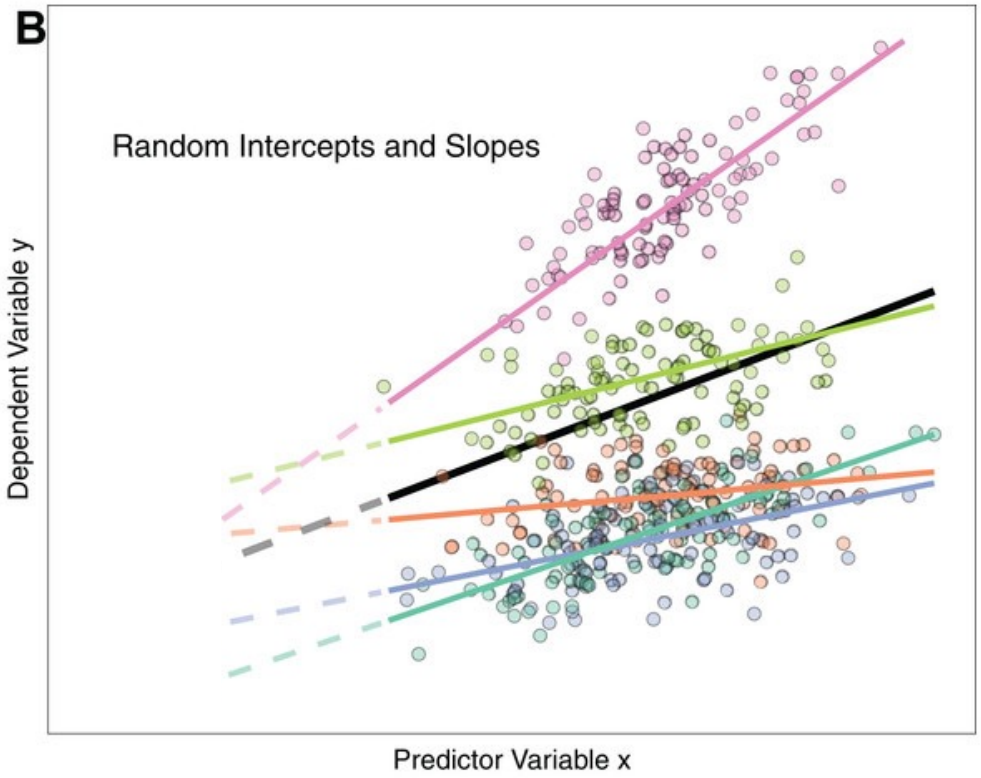
$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij}$$

with $j = 1, \ldots, n_i$, $i = 1, \ldots, n$.

$\Rightarrow$ each subject mean response varies not only in the baseline response mode ($t_{i1} = 0$) but also in thems of changes in their response over time

Indiduals may have lower or higher baseline reponse ($\beta_0 + b_{0i}$) than the population average (β_0); and steeper or less steep rate of change ($\beta_1 + b_{1i}$) than the population average (β_1)





The population mean is represented in black

Interpretation of the parameters

- Of course, the baseline response may not necessarily be $t = 0$ in your data. More in general, it could be a specified time $\bar{t}$.
- Then, for a generic individual at time t, the conditional model is $\mathrm{E}[Y|\boldsymbol{b}, t] = (\beta_0 + b_{0i}) + (\beta_1 + b_{i1})\,(t - \bar{t})$

  and the marginal model is $\mathrm{E}[Y|t] = \beta_0 + \beta_1(t - \bar{t})$

  so that β_0 is the expected (population averaged) response at $t = \bar{t}$ and the slope parameter β_1 is the expected (population averaged) change in response for a unit increase in time.
- An alternative interpretation is that β_1 is the change in response for a unit change in t for a "typical" individual, that is, an individual with $b_1 = 0$
- For a specific individual i, $\beta_0 + b_{0i}$ is the expected response at $t = \bar{t}$, and $\beta_1 + b_{1i}$ is the expected change in response for a unit increase in time.
- (Section 8.4.3 may be a bit confusing)

Remark 1.

- The effects of covariates (e.g., due to treatments, exposures, or background characteristics of the individuals) can be incorporated by allowing the means of the intercepts and slopes to depend on these covariates (e.g., by allowing them to vary across the different treatment groups or levels of exposure).

- For example, consider an hypothetical two-group study comparing a treatment and a control group. If the mean response changes in an approximately linear fashion over time, but with the means of the intercepts and slopes depending on group, the following linear mixed effects model can be adopted:

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + \beta_3 \,\text{Group}_i + \beta_4 t_{ij} \times \text{Group}_i + b_{1i} + b_{2i} t_{ij} + \epsilon_{ij}$$

where Group $_i = 1$ if the i^{th} individual was assigned to the treatment, and Group $_i = 0$ otherwise.

Remark 1 - ctd.

- In this model the design matrix X_i has the following form for the control group:

$$X_i = \begin{pmatrix} 1 & t_{i1} & 0 & 0 \\ 1 & t_{i2} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & 0 & 0 \end{pmatrix}$$

whereas for the treatment group the design matrix is given by

$$X_i = \begin{pmatrix} 1 & t_{i1} & 1 & t_{i1} \\ 1 & t_{i2} & 1 & t_{i2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{in_i} & 1 & t_{in_i} \end{pmatrix}$$

Remark 1 - ctd.

- Note that the design matrix Z_i has the same form for both the treatment and control groups,

$$Z_i = \begin{pmatrix} 1 & t_{11} \\ 1 & t_{i2} \\ \vdots & \vdots \\ 1 & t_{in_1} \end{pmatrix}$$

Remark 2

- Next, consider the covariance among the components of Y_i in the linear mixed effects model with randomly varying intercepts and slopes. Let's consider an unstructured covariance matrix for the random effects, i.e.

$$\text{Var}\left(b_{1j}\right) = g_{11}, \text{Var}\left(b_{2i}\right) = g_{22}, \text{Cov}\left(b_{1i}, b_{2i}\right) = g_{12}$$

- Then,

$$\begin{aligned}
\text{Var}\left(Y_{ij}\right) &= \text{Var}\left(X'_{ij}\beta + Z'_{ij}b_i + \epsilon_{ij}\right) \\
&= \text{Var}\left(Z'_{ij}b_i + \epsilon_{ij}\right) \\
&= \text{Var}\left(b_{1i} + b_{2i}t_{ij} + \epsilon_{ij}\right) \\
&= \text{Var}\left(b_{1i}\right) + 2t_{ij}\,\text{Cov}\left(b_{1i}, b_{2i}\right) + t_{ij}^2\,\text{Var}\left(b_{2i}\right) + \text{Var}\left(\epsilon_{ij}\right) \\
&= g_{11} + 2t_{ij}g_{12} + t_{ij}^2 g_{22} + \sigma^2
\end{aligned}$$

Remark 2 - ctd.

- Similarly it can be shown that:

$$\begin{aligned}
\text{Cov}\left(Y_{ij}, Y_{ik}\right) &= \text{Cov}\left(X'_{ij}\beta + Z'_{ij}b_i + \epsilon_{ij}, X'_{ik}\beta + Z'_{ik}b_i + \epsilon_{ik}\right) \\
&= \text{Cov}\left(Z'_{ij}b_i + \epsilon_{ij}, Z'_{ik}b_i + \epsilon_{ik}\right) \\
&= \text{Cov}\left(b_{1i} + b_{2i}t_{ij} + \epsilon_{ij}, b_{1i} + b_{2i}t_{ik} + \epsilon_{ik}\right) \\
&= \text{Var}\left(b_{1i}\right) + \left(t_{ij} + t_{ik}\right)\text{Cov}\left(b_{1i}, b_{2i}\right) + t_{ij}t_{ik}\,\text{Var}\left(b_{2i}\right) \\
&= g_{11} + \left(t_{ij} + t_{ik}\right)g_{12} + t_{ij}t_{ik}g_{22}
\end{aligned}$$

- Thus in this model for longitudinal data the covariance matrix, $\text{Cov}\left(Y_i\right)$, can be expressed as a function of time, t_{ij}. In particular, with the inclusion of random intercepts and slopes, the variance can increase or decrease over time as a quadratic function of the times of measurement.
- The expression for $\text{Var}\left(Y_{ij}\right)$ implies that the variance increases over time (for $t_{ij} \geq 0$) when $\text{Cov}\left(b_{1i}, b_{2i}\right) \geq 0$ but can decrease over time when $\text{Cov}\left(b_{1i}, b_{2i}\right) < 0$. Similarly the magnitude of the covariance (and correlation) between a pair of responses, say Y_{ij} and Y_{ik}, depends on the time separation between them (t_{ij} and t_{ik}).

General ANOVA model

- A general ANOVA model is defined as

$$y_{ij} = \mu + \boldsymbol{Z}\boldsymbol{b}_i + \epsilon_{ij}$$

where

$$\boldsymbol{Z}\boldsymbol{b}_i = Z_{1i}\, b_{1i} + \cdots + Z_{is}\, b_{is}$$

- We assume that $b_{ij} \overset{i.i.d.}{\sim} N(0, \sigma_j^2)$ and $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \tau^2)$ and the b_i's and ϵ_{ij}'s are independent
- For ANOVA models, a natural set of variance components ($\boldsymbol{\alpha}$) is $\tau^2, \sigma_1^2, \ldots, \sigma_s^2$. We call this form of variance components the original form. Alternatively, the Hartley-Rao form of variance components (Hartley and Rao, 1967) is:
$\lambda = \tau^2, \gamma_1 = \sigma_1^2/\tau^2, \ldots, \gamma_s = \sigma_s^2/\tau^2$

General ANOVA model

- If we marginalize the random vectors, the general ANOVA model can be described as

$$y \sim N(\mu, V)$$

with $V = \tau^2 I_n + \sum_{i=1}^{s} \sigma \boldsymbol{Z}_i \boldsymbol{Z}_i'$, which again characterizes the dependent correlation structure induced by the random effects.

Baseline and Centering

Baseline and Centering

- Note that the covariance matrix for the vector of random effects is not invariant to a linear transformation of Z_i.
- Linear transformations of the columns of Z_i alter the interpretation of b_i and change the estimates of the variances and covariances of the random effects.

- For example, in the linear mixed effects model with randomly varying intercepts and slopes, centering of the times of measurement (e.g., $t_{ij} - \bar{t}$, for $\bar{t} \neq 0$) alters the interpretation of the intercepts, and this leads to a change in the estimated variance of the random intercepts and their covariance with the random slopes.

- For example, in the model with untransformed times of measurement the variance of the "intercepts" is a measure of the between-subject variability in the response at time zero. However, in the model with transformed times of measurement, say centered at $\bar{t} \neq 0$, the variance of the "intercepts" is a measure of the between-subject variability in the response at time $\bar{t}$.

Baseline and Centering - ctd

- Linear transformations of components of Z_i produce equivalent mixed effects models only when the covariance matrix, $\boldsymbol{D}$, has been left unstructured. When $\boldsymbol{D}$ is unstructured the appropriate changes to the variances and covariances of the random effects can be produced.

- In addition, centering should be done with care:
- t_{ij} $\longrightarrow$ time from baseline (enrollment) $\longrightarrow$ OK!
- t_{ij} $\longrightarrow$ age $\longrightarrow$ be careful!
  - centering at an individual's age at the j-th measurement occasion (e.g. their mean over occasions) $\longrightarrow$ doesnt' make sense
- Rule of thumb: center the times of measurement for all subjects at some common fixed age within the age range of the study participants (i.e., $t_{ij} - a$, for some fixed value a, e.g. time of menarche) (preferable)
- By centering at a common value, the intercept is interpretable as the mean response at that common value for time (or age) and Var (b_{1i}) also has a meaningful interpretation.

Covariance Models for Clustered Data

- Let us consider again the general formulation:

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

- In the previous models, we have seen that a common assumption is to assume

$$\boldsymbol{b}_i \sim_{iid} \mathbf{N}_{q+1}(\mathbf{0}, \boldsymbol{D}) \quad \text{and} \quad \boldsymbol{\epsilon}_i \sim_{ind} \mathbf{N}_{n_i}(\mathbf{0}, \boldsymbol{E}_i)$$

- Also, it is common to assume uncorrelated errors, giving the simplified form $\boldsymbol{E}_i = \sigma^2_\epsilon \mathbf{I}_{n_i}$
- We will refer to σ^2_ϵ as the measurement error variance
- The *marginal* variances and covariances can be directly assessed from the observed data (later more on this), so we will look at their form

Exchangeable or compound symmetry

- We first consider the random intercepts only model $\boldsymbol{z}_i \boldsymbol{b}_i = \mathbf{1}_{n_i} b_i$ with $\text{var}\,(b_i) = \sigma_0^2$, along with $\boldsymbol{E}_i = \sigma_\epsilon^2 \mathbf{I}_{n_i}$
- Then,

$$\text{var}\,(\boldsymbol{Y}_i) = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \dots & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \rho & \rho & 1 & \dots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \dots & 1 \end{bmatrix}$$

  where $\sigma^2 = \sigma_\epsilon^2 + \sigma_0^2$ and $\rho = \sigma_0^2 / \sigma^2$.
- The entire covariance function is reduced to two variance parameters $\boldsymbol{\alpha} = \left[\sigma_\epsilon^2, \sigma_0^2\right]$.

Exchangeable or compound symmetry

- Since the covariance structure is summarized by only two parameters, the exchangeable or compound symmetry is not that flexible; for example, it postulates a constant correlation across occasions, whereas correlation is typically expected to decrease across occasions.
- Hence, the exchangeable model is particularly appropriate for clustered data with no time ordering as may arise, for example, in a split-plot design, or for multiple measurements within a family.
- For longitudinal data, only over short time scales

Serial Correlation

- An extension to the two-stage LME discussed earlier considers:

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\delta}_i + \boldsymbol{\epsilon}_i$$

  with
  - $\boldsymbol{b}_i$ representing individual-specific random effects;
  - $\boldsymbol{\delta}_i$ representing serial dependence;
  - $\boldsymbol{e}_i$ representing measurement error

Serial Correlation

- We assume

$$
\begin{aligned}
\mathrm{E}\left[\epsilon_i\right] &= \mathbf{0}, \mathrm{var}\left(\epsilon_i\right) = \sigma_\epsilon^2 \mathbf{I}_{n_i} \\
\mathrm{E}\left[\boldsymbol{b}_i\right] &= \mathbf{0}, \quad \mathrm{var}\left(\boldsymbol{b}_i\right) = \boldsymbol{D} \\
\mathrm{E}\left[\boldsymbol{\delta}_i\right] &= \mathbf{0}, \mathrm{var}\left(\boldsymbol{\delta}_i\right) = \sigma_\delta^2 \boldsymbol{R}_i \\
\mathrm{cov}\left(\boldsymbol{b}_i, \boldsymbol{\epsilon}_{i'}\right) &= \mathbf{0}, \quad i, i' = 1, \ldots, m \\
\mathrm{cov}\left(\boldsymbol{b}_i, \boldsymbol{\delta}_{i'}\right) &= \mathbf{0}, i, i' = 1, \ldots, m \\
\mathrm{cov}\left(\boldsymbol{\delta}_i, \boldsymbol{\epsilon}_{i'}\right) &= \mathbf{0}, i, i' = 1, \ldots, m \\
\mathrm{cov}\left(\boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_{i'}\right) &= \mathbf{0} \quad i \neq i' \\
\mathrm{cov}\left(\boldsymbol{\delta}_i, \boldsymbol{\delta}_{i'}\right) &= \mathbf{0} \quad i \neq i' \\
\mathrm{cov}\left(\boldsymbol{b_i}, \boldsymbol{b_{i'}}\right) &= \mathbf{0} \quad i \neq i'
\end{aligned}
$$

Auto-regressive structure

- We can induce a wide variety of time-varying correlations between observations by appropriately modeling the correlation matrix $\boldsymbol{R}_i$.

- For example, we can consider a AR(1) process:

$$\delta_{ij} = \rho \delta_{i,j-1} + u_{ij}$$

with $\mathrm{E}[\boldsymbol{u}_i] = \mathbf{0}, \mathrm{var}(\boldsymbol{u}_i) = \sigma_u^2 \mathbf{I}_{n_i}$.

- It is easy to show that in this case

$$\boldsymbol{R}_i = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n_i-1} \\ \rho & 1 & \rho & \dots & \rho^{n_i-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n_i-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n_i-1} & \rho^{n_i-2} & \rho^{n_i-3} & \dots & 1 \end{bmatrix}$$

in the case of equally spaced times.

For example, to specify an $AR(1)$ correlation structure with $\phi = 0.8$, position variable given by observation order within-group, and grouping variable `Subject`, we use

```
> cs1AR1 <- corAR1( 0.8, form = ~ 1 | Subject )
> cs1AR1 <- initialize( cs1AR1, data = Orthodont )
> corMatrix( cs1AR1 )
$M01:
      [,1] [,2] [,3]  [,4]
[1,] 1.000 0.80 0.64 0.512
[2,] 0.800 1.00 0.80 0.640
[3,] 0.640 0.80 1.00 0.800
[4,] 0.512 0.64 0.80 1.000
```

As described in §5.3.1, the $AR(1)$ model is equivalent to an $ARMA(1,0)$ model, so that the corARMA class can also be used to represent an corAR1 object. However, the corAR1 methods are designed to take advantage of the particular structure of the $AR(1)$ model, and are substantially more efficient than the corresponding corARMA methods.

Pinheiro & Bates (2000), Mixed Effects Models in S- and S-Plus, Springer

AR(1) covariance function

- However, $\boldsymbol{R}_i$ has a Toeplitz structure:

$$\text{var}\left(\boldsymbol{\delta}_i\right) = \sigma_\delta^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n_i-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n_i-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n_i-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n_i-1} & \rho_{n_i-2} & \rho_{n_i-3} & \cdots & 1 \end{bmatrix}$$

which is most appropriate when the measurements are made at equal intemals of time and does not take into account distances between measurements

Exponential Covariance Function

- The model can be extended to unequally spaced times to give covariance

$$\text{cov}\left(\delta_{ij}, \delta_{ik}\right) = \sigma_{\delta}^{2} \rho^{|t_{ij} - t_{ik}|}$$

- Correlations decline over time as the separation between pairs ef repeared measures increases

- Still, the decay is quite fast with respect to what one sees in LDA data

Nested Covariance Structures

Unstructured Covariance Matrix

- An unstructured covariance structure allows for different variances at each occasion $\sigma^2_{\delta 1}, \ldots, \sigma^2_{\delta n_i}$ and distinct correlations for each pair of responses, that is,

$$\text{corr}\left(\boldsymbol{\delta}_i\right) = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1n_i} \\ \rho_{21} & 1 & \rho_{23} & \cdots & \rho_{2n_i} \\ \rho_{31} & \rho_{32} & 1 & \cdots & \rho_{3n_i} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n_i1} & \rho_{n_i2} & \rho_{n_i3} & \cdots & 1 \end{bmatrix}$$

  with $\rho_{jk} = \rho_{kj}$, for $j, k = 1, \ldots, n_i$.

- This model contains $n_i\left(n_i + 1\right)/2$ parameters per individual, which is a large number if n_i is large.

  If one has a common design across individuals, it may be plausible to fit this model, but one would still need a large number of individuals m, in order for inference to be reliable.

  As usual, there is a trade-off between flexibility and parsimony.

Choice among covariance pattern models

- We will discuss diagnostic techniques later on, but of course one issue is how to choose the "most appropriate" covariance model and how misspecification is going to affect the inference.

Choice among covariance pattern models

- We will discuss diagnostic techniques later on, but of course one issue is how to choose the "most appropriate" covariance model and how misspecification is going to affect the inference.
- The model for the covariance depends on the choice of a model for the mean
- General guidelines: one may want to find a maximal model to the mean, then decide on the covariance:
  - easy if there are not many variables/interactions
  - based on subject-matter considerations
- Given a maximal model for the mean, a sequence of covariance pattern models can be fit to the data

Choice among covariance pattern models

- Any model for the covariance depends on the assumed model for the mean, since a model for the covariance tries to account for the covariance among residuals

$$(y_{is} - \mu_{is}(\beta))' (y_{ik} - \mu_{ik}(\beta))$$

- Maximal models typically include treatment covariates (RCT) or quasi-treatment covariates (exposine groups in observational studies) and their interaction with time
- The choice of wether to include additional variables/interactions must be made on subject-matter grounds, not through an automatic procedure!!
- We will see soon that when pair of models are nested (e.g., the compound symmetry model is nested within the Toeplitz), a LRT (REML) can be constructed to compare a full and reduced model for the covariance

Structure	*Example*	*Parameters*
Compound symmetry (CS)	$\sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho & \rho \\ & 1 & \rho & \rho & \rho \\ & & 1 & \rho & \rho \\ & & & 1 & \rho \\ & & & & 1 \end{bmatrix}$	2
First-order autoregressive [AR(1)]	$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ & 1 & \rho & \rho^2 & \rho^3 \\ & & 1 & \rho & \rho^2 \\ & & & 1 & \rho \\ & & & & 1 \end{bmatrix}$	2
Unstructured (UN)	$\begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ & \sigma_{22}^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ & & \sigma_{33}^2 & \sigma_{34} & \sigma_{35} \\ & & & \sigma_{44}^2 & \sigma_{45} \\ & & & & \sigma_{55}^2 \end{bmatrix}$	15

NOTE: All matrices are symmetric, so only their upper triangles are given. Greek letters represent unknown parameters. The parameter ρ satisfies $|\rho| < 1$.

The Three Most Common Covariance Structures

Heterogeneous compound symmetry (CSH)

$$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \sigma_1\sigma_4\rho & \sigma_1\sigma_5\rho \\ & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho & \sigma_2\sigma_5\rho \\ & & \sigma_3^2 & \sigma_3\sigma_4\rho & \sigma_3\sigma_5\rho \\ & & & \sigma_4^2 & \sigma_4\sigma_5\rho \\ & & & & \sigma_5^2 \end{bmatrix}$$

Heterogeneous first-order autoregressive [ARH(1)]

$$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 & \sigma_1\sigma_5\rho^4 \\ & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 & \sigma_2\sigma_5\rho^3 \\ & & \sigma_3^2 & \sigma_3\sigma_4\rho & \sigma_3\sigma_5\rho^2 \\ & & & \sigma_4^2 & \sigma_4\sigma_5\rho \\ & & & & \sigma_5^2 \end{bmatrix}$$

Heterogeneous Toeplitz (TOEPH)

$$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_2 & \sigma_1\sigma_4\rho_3 & \sigma_1\sigma_5\rho_4 \\ & \sigma_2^2 & \sigma_2\sigma_3\rho_1 & \sigma_2\sigma_4\rho_2 & \sigma_2\sigma_5\rho_3 \\ & & \sigma_3^2 & \sigma_3\sigma_4\rho_1 & \sigma_3\sigma_5\rho_2 \\ & & & \sigma_4^2 & \sigma_4\sigma_5\rho_1 \\ & & & & \sigma_5^2 \end{bmatrix}$$

Statistical Methods for Correlated Data

Likelihood Inference for Linear Mixed Models

Michele Guindani

Department of Biostatistics
UCLA

Inference for LMMs

- We consider methods for inference in LMMs:

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

  Standard methods of estimation in linear (better, gaussian) mixed models are maximum likelihood (ML) and restricted maximum likelihood (REML) methods

Inference for LMMs

- We consider methods for inference in LMMs:

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} + \boldsymbol{z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

  Standard methods of estimation in linear (better, gaussian) mixed models are maximum likelihood (ML) and restricted maximum likelihood (REML) methods

- In order to conduct likelihood inference, we need distributional assumptions on the random and measurement error terms. As discussed previously, we assume

$$\epsilon_i|\sigma^2_\epsilon \sim_{iid} \mathbf{N}_{n_i}(\mathbf{0}, E_i) \quad \text{and} \quad \boldsymbol{b}_i|\boldsymbol{D} \sim_{iid} \mathbf{N}_{q+1}(\mathbf{0}, \boldsymbol{D})$$

  and we further assume $E_i = \sigma^2_\epsilon \mathbf{I}_{n_i}$ and D *unstructured*:

$$\boldsymbol{D} = \begin{bmatrix} \sigma^2_0 & \sigma_{01} & \sigma_{02} & \dots & \sigma_{0q} \\ \sigma_{10} & \sigma^2_1 & \sigma_{12} & \dots & \sigma_{1q} \\ \sigma_{20} & \sigma_{21} & \sigma^2_2 & \dots & \sigma_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{q0} & \sigma_{q1} & \sigma_{q2} & \dots & \sigma^2_q \end{bmatrix}$$

Inference for LMMs

- Hence, the vector of variance-covariance parameters is $\boldsymbol{\alpha} = \left[\sigma^2_\epsilon, \boldsymbol{D}\right]$
- In the frequentist domain, the random part component is used to model within-individual correlations, serial correlations, etc.
- Frequentist inferences usually focuses primarily on the fixed effects regression parameters $\boldsymbol{\beta}$ and the variance components α, although we could also potentially be interested on the random effects $\boldsymbol{b} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m]^{\mathrm{T}}$.
- That is, we are often interested in the marginal and mean variances:

$$\mathrm{E}\left[\boldsymbol{Y}_i|\boldsymbol{\beta}\right] = \boldsymbol{\mu}_i(\boldsymbol{\beta}) = \boldsymbol{x}_i\boldsymbol{\beta}$$

$$\mathrm{var}\left(\boldsymbol{Y}_i|\boldsymbol{\alpha}\right) = \boldsymbol{V}_i(\boldsymbol{\alpha}) = \boldsymbol{z}_i\boldsymbol{D}\boldsymbol{z}_i^{\mathrm{T}} + \sigma^2_\epsilon\mathbf{I}_{n_i}$$

Inference of LMMs

- In order to conduct inference on the fixed effects $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, we need to integrate over the random effects in the two-stage model:

$$p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\alpha}) = \int_{\mathcal{S}_{\boldsymbol{b}}} p(\boldsymbol{y}|\boldsymbol{b},\boldsymbol{\beta},\boldsymbol{\alpha}) \times p(\boldsymbol{b}|\boldsymbol{\beta},\boldsymbol{\alpha}) d\boldsymbol{b}$$

- Due to the conditional independencies in the likelihood

$$p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\alpha}) = \prod_{i=1}^{m} \int_{\mathcal{S}_{\boldsymbol{b}_i}} p\left(\boldsymbol{y}_i|\boldsymbol{b}_i,\boldsymbol{\beta},\sigma_\epsilon^2\right) \times p\left(\boldsymbol{b}_i|\boldsymbol{D}\right) d\boldsymbol{b}_i$$

and since a convolution of normal distributions is still normal, we have

$$\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{\alpha} \sim \mathbf{N}_{n_i}\left[\boldsymbol{\mu}_i(\boldsymbol{\beta}), \boldsymbol{V}_i(\boldsymbol{\alpha})\right]$$

.

- The log-likelihood is

$$(\boldsymbol{\beta}, \boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{m} \log |\boldsymbol{V}_i(\boldsymbol{\alpha})| - \frac{1}{2}\sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{V}_i(\boldsymbol{\alpha})^{-1} (\boldsymbol{y}_i - \boldsymbol{x}_i\boldsymbol{\beta})$$

- We need to maximize the previous expression with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. We obtain the score functions:

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{Y}_i - \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} \boldsymbol{x}_i \boldsymbol{\beta} \\
&= \sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{x}_i \boldsymbol{\beta}) \\
\frac{\partial l}{\partial \alpha_r} &= \frac{1}{2} \left\{ (y - \boldsymbol{x}_i\beta)' \boldsymbol{V}_i^{-1} \frac{\partial \boldsymbol{V}_i}{\partial \alpha_r} \boldsymbol{V}_i^{-1} (y - \boldsymbol{x}_i\beta) - \mathrm{tr}\left(\boldsymbol{V}_i^{-1} \frac{\partial \boldsymbol{V}_i}{\partial \alpha_r} \right) \right\}
\end{aligned}$$

$r = 1, \ldots, q$

- Hence, we obtain the MLE for $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})^{-1} \boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})^{-1} \boldsymbol{y}_i \right)$$

  which is a generalized least squares estimator (GLS).

- Hence, we obtain the MLE for $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})^{-1} \boldsymbol{x}_i \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i(\widehat{\boldsymbol{\alpha}})^{-1} \boldsymbol{y}_i \right)$$

  which is a generalized least squares estimator (GLS).

- $\boldsymbol{V}_i(\hat{\boldsymbol{\alpha}})$ is a function of an estimate $\hat{\alpha}$ of the variance components (plug-in).

- If $D = \mathbf{0}$, then $\boldsymbol{V} = \sigma_\epsilon^2 \mathbf{I}_N$, $N = \sum_{i=1}^{m} n_i$, and $\widehat{\boldsymbol{\beta}}$ corresponds to the ordinary least squares estimator

- It is easy to prove that

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m} \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{V}_i(\boldsymbol{\alpha})^{-1} \boldsymbol{x}_i \right)^{-1}$$

Some remarks

- The result can be seen as the maximization of a profile likelihood of $\boldsymbol{\beta}$ given $V(\boldsymbol{\alpha})$ or an estimate thereof.
- $\boldsymbol{V}_i(\hat{\boldsymbol{\alpha}}) = \boldsymbol{Z}_i D(\hat{\boldsymbol{\alpha}}) \boldsymbol{Z}_i' + \boldsymbol{E}_i$
- The expected information matrix is block diagonal:

$$\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \begin{bmatrix} \boldsymbol{I}_{\beta\beta} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{\alpha\alpha} \end{bmatrix}$$

  so there is asymptotic independence between $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ and any consistent estimator of α will give an asymptotically efficient estimator for β
- The estimator $\widehat{\boldsymbol{\beta}}$ is linear in the data Y_i, and so under normality of the data, $\hat{\boldsymbol{\beta}}$ is normal also.

- Under correct specification of the variance model, and with a consistent estimator $\widehat{\boldsymbol{\alpha}}$

$$\left(\sum_{i=1}^{m} x_i V_i(\widehat{\alpha})^{-1} x_i\right)^{1/2} \left(\widehat{\beta}_m - \beta\right) \to_d \mathrm{N}_{k+1}(0, \mathrm{I})$$

  as $m \to \infty$.
- Since $\widehat{\boldsymbol{\beta}}$ is linear in $\boldsymbol{Y}$, These properties are valid for large samples even if the sampling distribution of Y_i is not multivariate normal (Laird and Ware, 1982).
- The second moments of the data need to be correctly specified.

Optimization

- The detailed maximization process in linear mixed models follows the standard procedures applied for general linear models.
- More specifically, optimization of the profiled log-likelihood is usually accomplished through EM iterations or through Newton-Raphson iterations (Laird and Ware, 1982)
- The EM algorithm (Dempster, Laird and Rubin, 1977) is a popular iterative algorithm for likelihood estimation in models with incomplete data.
- The EM iterations for the LME model are based on regarding the random effects, such as the $\boldsymbol{b}_i, i = 1, \ldots, M$, as unobserved data.
- Each iteration of the EM algorithm results in an increase in the log-likelihood, till convergence (stats 230)

EM algorithm - idea

- Starting from parameter guesses, $\boldsymbol{\alpha}, \boldsymbol{\beta}$, the following steps are iterated:
  1. Find the distribution of $\boldsymbol{b}|\boldsymbol{y}$ according to the current parameter estimates (we will see soon how to compute $\boldsymbol{b}|\boldsymbol{y}$)
  2. Treating the distribution from 1 as fixed (rather than depending on $\boldsymbol{\alpha}, \boldsymbol{\beta}$), find an expression for $Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathbb{E}_{|\boldsymbol{y}}\{\log f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})\}$ as a function of $\boldsymbol{\alpha}, \boldsymbol{\beta}$, using the distribution from 1. The $\mathbf{y}$ are treated as fixed, here. (This is the E-step.)
  3. Maximize the expression for $Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$ w.r.t. the parameters to obtain updated estimates $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$. (This is the M-step.)

EM algorithm - idea

- Note that the expectation in step 2 is taken with respect to the fixed distribution from step 1, which depends on the current parameter estimates. When evaluating $Q(\boldsymbol{\alpha}, \boldsymbol{\beta})$, we view $\log f(\mathbf{y}, \mathbf{b}|\boldsymbol{\beta})$ as a function of $\boldsymbol{\alpha}, \boldsymbol{\beta}$, but do not treat the distribution of $\mathbf{b}|\mathbf{y}$ as depending on these parameters.

Newton-Raphson

- The Newton-Raphson algorithm (Thisted, 1988) is one of the most widely used optimization procedures.
- It uses a first-order expansion of the score function (the gradient of the log-likelihood function) around the current estimate $\boldsymbol{\alpha}^{(w)}$ to produce the next estimate $\boldsymbol{\alpha}^{(w+1)}$.
- Each Newton-Raphson iteration requires the calculation of the score function and its derivative, the Hessian matrix of the log-likelihood.
- Under general conditions usually satisfied in practice, the Newton-Raphson algorithm converges quadratically.
- Because the calculation of the Hessian matrix at each iteration may be computationally expensive, simple, quicker to compute approximations are sometimes used, leading to the so-called Quasi-Newton algorithms.

Optimization

- Individual iterations of the EM algorithm are quickly and easily computed.
- Although the EM iterations generally bring the parameters into the region of the optimum very quickly, progress toward the optimum tends to be slow when near the optimum.
- Newton-Raphson iterations, on the other hand, are individually more computationally intensive than the EM iterations, and they can be quite unstable when far from the optimum. However, close to the optimum they converge very quickly
- A hybrid approach starts with an initial $\boldsymbol{\alpha}^{(0)}$, performing a moderate number of EM iterations, then switches to Newton-Raphson iterations.
- The lme function in the nlme package of R implements such a hybrid optimization scheme. It begins by calculating initial estimates of the $\boldsymbol{\alpha}$ parameters, then uses several EM iterations to get near the optimum, then switches toNewton-Raphson iterations to complete the convergence to the optimum. By default 25 EM iterations are performed before switching to Newton-Raphson iterations.

Testing Fixed effects

- Tests of fixed effects are typically done with either Wald or likelihood ratio (LRT) tests. With asymptotic distributions and independent predictors, Wald and LRT tests are equivalent.
- When a data set size is not large enough to be a good approximation of the asymptotic distribution or there is some correlation amongst the predictors, the Wald and LRT test results can vary considerably.
- Let $\tilde{\boldsymbol{L}}$ be a design vector or a design matrix of known weights for selected components in $\boldsymbol{\beta}$ and $\boldsymbol{L\beta}$ be a combination of interest.
- The sampling distribution of $\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}}$ is multivariate normal with mean $\boldsymbol{L\beta}$ and covariance matrix

$$\begin{aligned}\operatorname{cov}(\boldsymbol{L}\hat{\boldsymbol{\beta}}) &= \tilde{\boldsymbol{L}}\operatorname{cov}(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{L}}' \\ &= \tilde{\boldsymbol{L}}\left[\left(\sum_{i=1}^{N}\boldsymbol{X}_i'\hat{\boldsymbol{V}}_i^{-1}\boldsymbol{X}_i\right)^{-1}\right]\tilde{\boldsymbol{L}}'\end{aligned}$$

- Empirically, $\tilde{\boldsymbol{L}}$ is often designed to contain weight 1 to indicate the selected components in β or weight 0 for the components not selected.

Wald test

- The two hypotheses, $\mathrm{H}_0 : \tilde{\boldsymbol{L}}\boldsymbol{\beta} = 0$ versus $\mathrm{H}_\mathrm{A} : \tilde{\boldsymbol{L}}\boldsymbol{\beta} \neq 0$ can be tested by using the following Wald statistic:

$$W^2 = (\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}}) \left\{ \tilde{\boldsymbol{L}} \left[\left(\sum_{i=1}^{N} \boldsymbol{X}_i' \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i \right)^{-1} \right] \tilde{\boldsymbol{L}}' \right\}^{-1} (\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}})$$

where W^2 is the Wald statistic that asymptotically follows a chi-square distribution with rank($\tilde{\boldsymbol{L}}$) as the degrees of freedom.

Wald test

- The two hypotheses, $\mathrm{H}_0 : \tilde{\boldsymbol{L}}\boldsymbol{\beta} = 0$ versus $\mathrm{H}_\mathrm{A} : \tilde{\boldsymbol{L}}\boldsymbol{\beta} \neq 0$ can be tested by using the following Wald statistic:

$$W^2 = (\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}}) \left\{ \tilde{\boldsymbol{L}} \left[\left(\sum_{i=1}^{N} \boldsymbol{X}_i' \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i \right)^{-1} \right] \tilde{\boldsymbol{L}}' \right\}^{-1} (\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}})$$

  where W^2 is the Wald statistic that asymptotically follows a chi-square distribution with rank($\tilde{\boldsymbol{L}}$) as the degrees of freedom.

- Similarly, the approximate confidence interval, given α, is given by

$$\tilde{L}\hat{\beta} \pm t_{\mathrm{df},\alpha/2} \times \left\{ \tilde{\boldsymbol{L}} \left[\left(\sum_{i=1}^{N} \boldsymbol{X} \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i \right)^{-1} \right] \tilde{\boldsymbol{L}}' \right\}^{\frac{1}{2}}.$$

- For a test of a single component, a Z-test works as an approximate Wald test.

Bias of the Wald statistics

- The Wald statistics is considered to be biased downward because the variability in estimating the variance components is not considered (Dempster et al. 1981) in the ML estimates.
- It is perceived that this bias can be resolved by using approximate F-statistic about β.
- Let $\mathrm{H}_0 : \tilde{\boldsymbol{L}}\beta = 0$ versus $\mathrm{H}_\mathrm{A} : \tilde{\boldsymbol{L}}\beta \neq 0$, then we can use the F -statistic:

$$F = \frac{(\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}}) \left\{ \tilde{\boldsymbol{L}} \left[\left(\sum_{i=1}^{N} \boldsymbol{X}_i' \hat{\boldsymbol{V}}_i^{-1}(\hat{\alpha}) \boldsymbol{X}_i \right)^{-1} \right] \tilde{\boldsymbol{L}}' \right\}^{-1} (\tilde{\boldsymbol{L}}\hat{\boldsymbol{\beta}})}{\mathrm{rank}(\tilde{\boldsymbol{L}})}$$

where the degrees of freedom for the numerator is rank($\boldsymbol{L}$) and the degrees of freedom for the denominator needs to be estimated from the data. The uncertainty about the degrees of freedom for the denominator somewhat restricts the use of the F -test (several methods, omit).

Likelihood ratio test

- To test $\mathrm{H}_{0:}: \tilde{\boldsymbol{L}}\boldsymbol{\beta} = 0$ versus $\mathrm{H}_{\mathrm{A}:}: \tilde{\boldsymbol{L}}\beta \neq 0$ we could use the LRT.
- The likelihood ratio test compares the maximized log-likelihoods between two models, given by

$$G^2 = 2\log L\left(\hat{\theta}_{\text{full}}\right) - 2\log L\left(\hat{\theta}_{\text{reduced}}\right)$$

  where G^2 is the likelihood ratio statistic, $\log L\left(\hat{\theta}_{\text{reduced}}\right)$ is the log-likelihood function for the model without one or more parameters, and $\log L\left(\hat{\theta}_{\text{full}}\right)$ is the log-likelihood function containing all parameters.
- The likelihood ratio statistic is asymptotically distributed as χ^2 with the degrees of freedom being the difference in the number of fixed-effects parameters.
- If G^2 is associated with a p -value smaller than α, the null hypothesis about θ should be rejected.