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 from many books and available material on the topic.

- 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

8 Demidenko, E. (2013) Mixed Modes: Theory and Applications with R, Wiley

9 Davidian, M.(2017) Statistical Methods for Analysis With Missing Data - Class Notes

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?

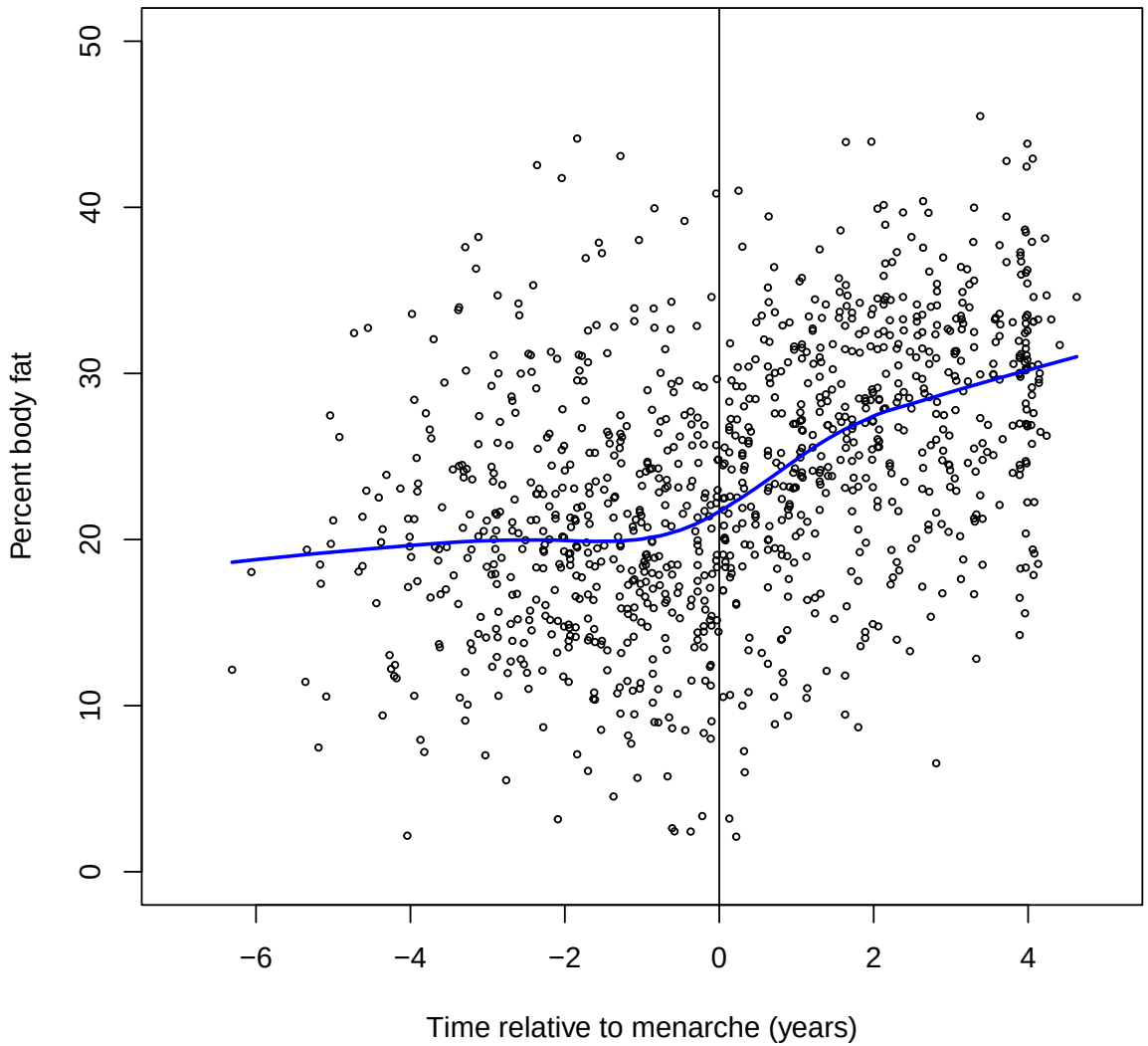
Example (MIT Growth and Development Study)

- 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



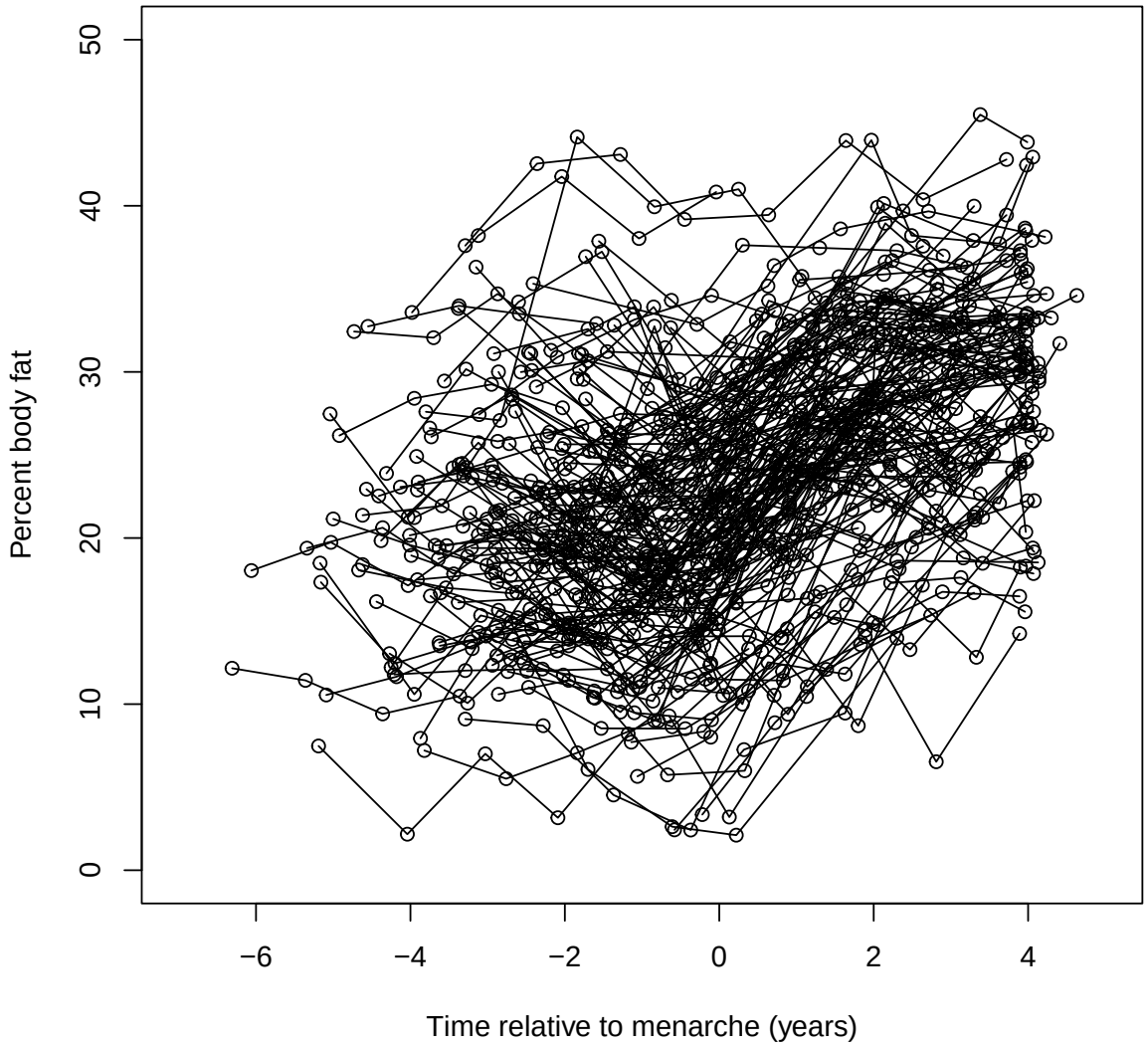
Naive solution

```
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

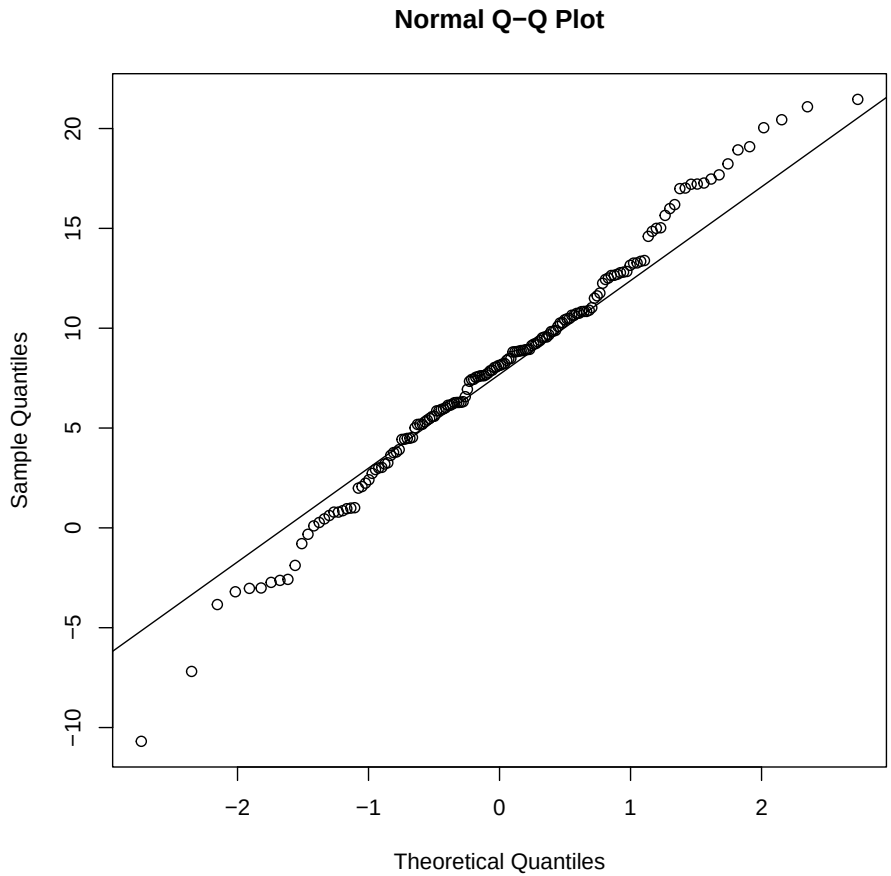
- Paired t-test - Look at each girl's first (pre-menarche) and last (post-menarche) PBF measurements and calculate the difference in PBF for each girl. Is there a significant positive mean gain in PBF?

```
# 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

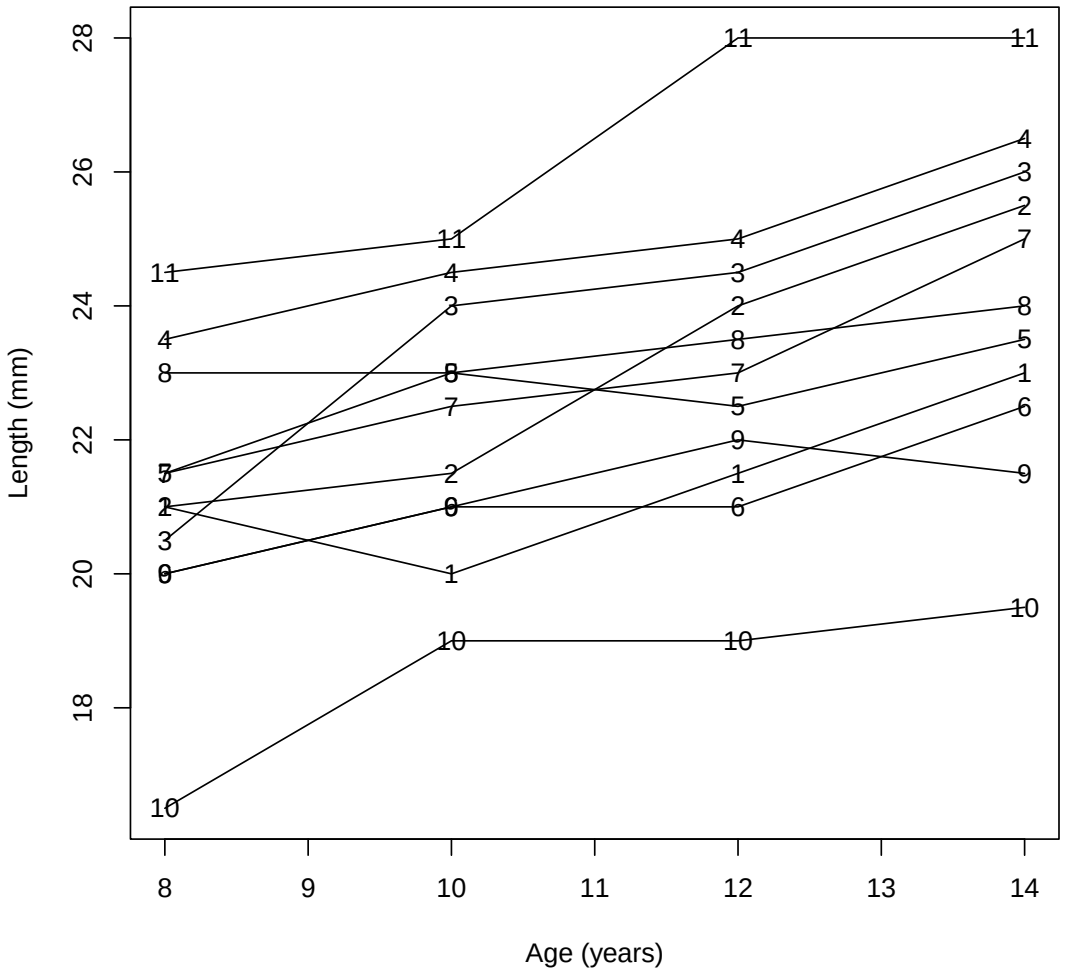
A second example: Dental Growth Curves

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:

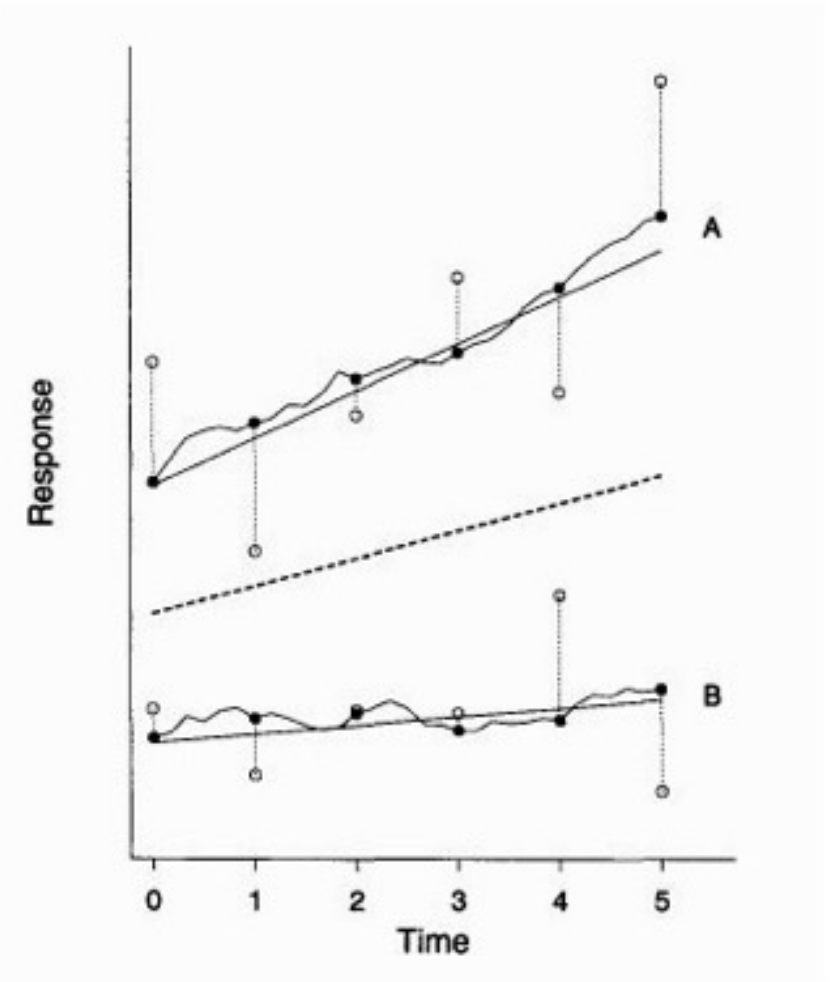
1. Between-individual heterogeneity
2. Within-individual biological variation
3. Measurement error

● 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.