Advanced Longitudinal Data Analysis

Sonja Greven

Johns Hopkins University, Spring 2009

Overview

1. Introduction
   - Longitudinal Data
   - Variation and Correlation
   - Different Approaches
2. Mixed Models
   - Linear Mixed Models
   - Generalized Linear Mixed Models
3. Marginal Models
4. Transition Models
5. Further Topics

Overview

1. Introduction
   - Longitudinal Data
   - Variation and Correlation
   - Different Approaches
2. Mixed Models
   - Linear Mixed Models
   - Generalized Linear Mixed Models
3. Marginal Models
4. Transition Models
5. Further Topics

A note about this class

Disclaimer:

- I will not discuss many examples of data analyses (although I will give illustrative examples and make some remarks on software).
- I will not cover exploratory tools. This does not mean, though, that these are not important! Find the best approach to analyze your data.

→ Francesca's LDA class

This course is complementary to Francesca's class and aims at giving you more of the underlying theory, and especially more details about estimation and inference in general.

Please feel to ask questions any time - especially during the more technical parts.

I asked Bruce and John to give short presentations on their current research involving longitudinal data. If you think your research in longitudinal data would also be interesting to share, please let me know!

Some Literature

Books

Diggle, P., P. Heagerty, K.-Y. Liang, and S. Zeger (2002).
Analysis of Longitudinal Data.
Oxford University Press.

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

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

More Papers and Books

Some papers and books are referenced in the slides. A bibliography of the most important ones will be on the website for further reading.

Acknowledgements

- Peter Diggle (Lancaster) and Amy Herring (UNC Chapel Hill) kindly shared their own LDA slides with me.
- Francesca Dominici and Ciprian Crainiceanu also shared their class material, so I could see what has been covered in other classes.

What is Longitudinal Data?

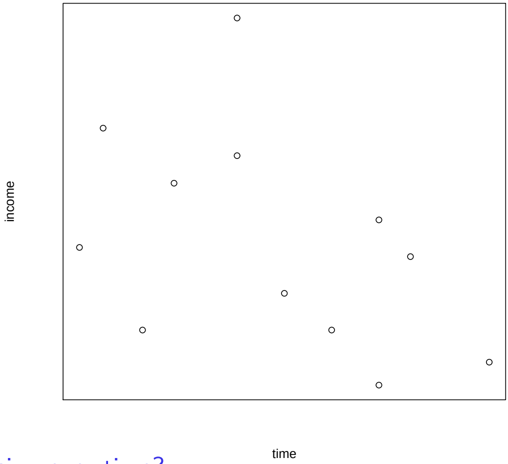
- I observational units (subjects, animals, . . .).
- n_i observations on the ith unit, $i = 1, \ldots, I$.
- The observations are taken at time points t_{ij}, $t_{i1} < \cdots < t_{in_i}$.

What is special about Longitudinal Data?

- Observations on the same subject/unit are not independent. They are (marginally) correlated.
- Observations have an ordering in time.
  (In contrast, for example, to clustered data, e.g. a litter of mice.)

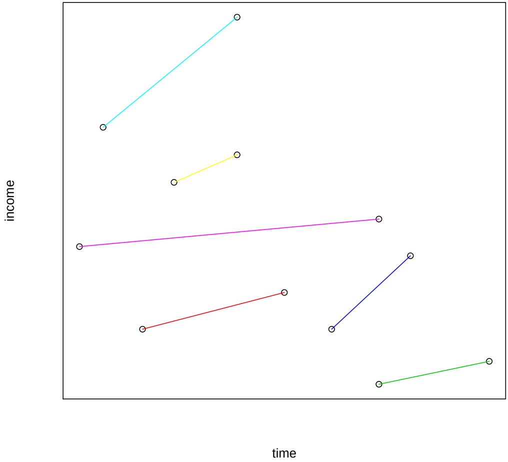
Advantages of Longitudinal Data

We can distinguish longitudinal from cross-sectional effects.



Is income decreasing over time?

Advantages of Longitudinal Data



Income is increasing over time for each person.
Starting salaries seem to be decreasing.

Advantages of Longitudinal Data

Cross-sectional (β_C) and longitudinal (β_L) effects

$$Y_{ij} = \beta_0 + \beta_C x_{i1} + \beta_L (x_{ij} - x_{i1}) + \varepsilon_{ij}$$

Without longitudinal information, we have to assume $\beta_C = \beta_L$. This is a strong assumption!
(E.g. $\beta_C =$ Increase in average starting salaries, $\beta_L =$ increase in salary after starting to work → opposite signs in our example)

Confounding

Can use each subject as its own control. But: confounding still possible by time-varying variables (e.g. seasonality, long-term trends).

Sources of Variation

We can distinguish different sources of variation

- Between subjects
- Within a subject over time
- Measurement error (at least if we have repeated measures at the same time)

Sources of Variation in Longitudinal Data

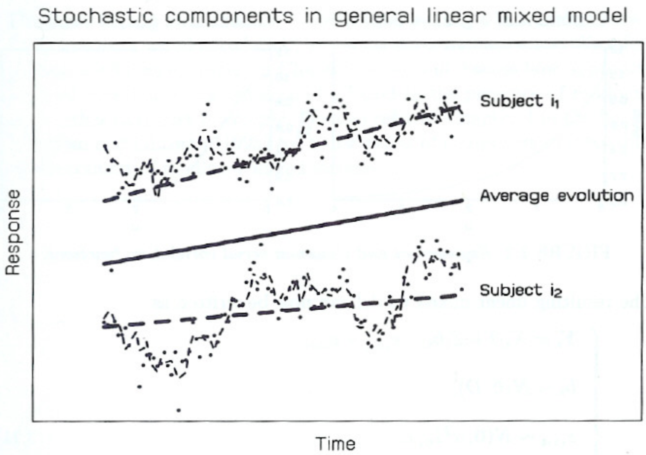


FIGURE 3.1. *Graphical representation of the three stochastic components in the general linear mixed model (3.11). The solid line represents the population-average evolution. The lines with long dashes show subject-specific evolutions for two subjects i_1 and i_2. The residual components of serial correlation and measurement error are indicated by short-dashed lines and dots, respectively.*

Source: Verbeke & Molenberghs (2000)

Sources of Variation in Longitudinal Data

Example: Hourly measurements of a bloodmarker:

- **Differences between people**:
  In average level and in average evolution in time. → b_0, b_1
- **Within a person over time**:
  Serial correlation due to e.g. long half-life of blood-marker, longer-term influences (alcohol, . . .) etc. → ε_{ij}
- **Measurement Error**:
  On top of that, we will probably not have exact measurements of the bloodmarker at time t. → ϵ_{ij}

Possible model:

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

$(b_{0i}, b_{1i})'$ bivariate normal, ε_{ij} auto-correlated, ϵ_{ij} i.i.d. error.
Often, we cannot estimate all of these well (especially ε_{ij} and ϵ_{ij}).

The (Semi-)Variogram

Usually, the natural setting for longitudinal data is continuous time. We can then view the $Y_{ij}, j = 1, \ldots, n_i$ as samples $Y_{ij} = Y_i(t_{ij})$ from I independent copies of an underlying continuous-time stochastic process $\{Y(t), t \in \mathbb{R}\}$.

An important tool to characterize $\{Y(t)\}$ is the variogram, which can be used to describe the covariance function,

$$\gamma(u) = \frac{1}{2} \mathrm{E}[\{Y(t) - Y(t-u)\}^2], \; u \geq 0.$$

It is well-defined for stationary and some non-stationary processes. For a stationary process (i.e. the joint probability distribution does not change when shifted in time), with $\mu = \mathrm{E}\{Y(t)\}$, $\sigma^2 = \mathrm{Var}\{Y(t)\}$ and $\rho(u) = \mathrm{Corr}\{Y(t), Y(t-u)\}$,

$$\begin{aligned}
\gamma(u) &= \frac{1}{2}\mathrm{E}[Y(t)^2 + Y(t-u)^2 - 2Y(t)Y(t-u)] \\
&= \frac{1}{2}\mathrm{E}[\{Y(t)-\mu\}^2 + \{Y(t-u)-\mu\}^2 - 2\{Y(t)-\mu\}\{Y(t-u)-\mu\}] \\
&= \frac{1}{2}[\sigma^2 + \sigma^2 - 2\sigma^2\rho(u)] = \sigma^2\{1-\rho(u)\}.
\end{aligned}$$

The (Semi-)Variogram

The empirical variogram is calculated using observed half-squared-differences between pairs of residuals (→ stationarity)

$$v_{ijk} = \frac{1}{2}(r_{ij} - r_{ik})^2,$$

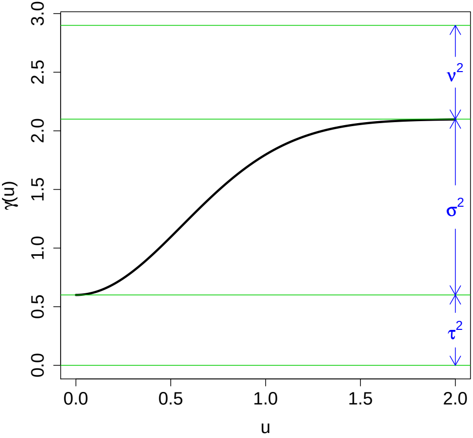
e.g. binned according to corresponding time-differences $u_{ijk} = t_{ij} - t_{ik}$.

The (Semi-)Variogram

Example: Random intercept plus serial correlation plus measurement error.

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \delta_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + \varepsilon_{ij} + \epsilon_{ij},$$

$b_{0i} \overset{iid}{\sim} (0, \nu^2)$, ε_{ij} auto-correlated with variance σ^2, correlation $\rho(u)$, $\epsilon_{ij} \overset{iid}{\sim} (0, \tau^2)$.



Variogram for δ
$\gamma(u) = \underbrace{\tau^2}_{nugget} + \sigma^2\{1 - \rho(u)\}$

$\text{Var}(\delta_{ij}) = \nu^2 + \sigma^2 + \tau^2$
$> \lim_{u\to\infty} \gamma(u) = \tau^2 + \sigma^2$

and

$\lim_{u\to 0} \gamma(u) = \tau^2 > 0$

→ information on all three components from repeated measurements

Why is it Important to Take the Correlation into Account?

If we ignore the correlation in our data:

- Our standard errors will be incorrect and our inference thus invalid.
- We lose efficiency in estimating the mean parameters.
- We have sub-optimal protection against biases caused by missing data.

Also: Often the correlation and/or decomposition of the variance is actually of scientific interest.

Example: New bloodmarker - how large is

- measurement error → how precisely can the lab measure?
- intra-subject variability → is one measurement representative of the subject's average marker level?
- inter-subject variability → is a general reference value useful in detecting individual unusual results?

Different Viewpoints of Correlation

- **Marginal Models**: Marginally, observations are correlated. We can model this and/or account for it with robust standard errors (GEE).
- **Mixed Models**: Observations are correlated, because they are from the same subject and share the same underlying processes. Conditional on these, observations are independent.
  (Or residual correlation left, then additionally model that.)
- **Transition/Markov Models**: Observations are correlated, because the past influences the presence.
  (Typical here: Past = last q observations → Markov property.)

We can model correlation or try to explain it, depending on our objectives.
Shift: Correlation → mean.

The Three Approaches for the Linear Model

Consider a simple linear regression model (e.g. for infant growth, Y = height)

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \epsilon_{ij}.$$

Marginal Model: Specifies a model for mean (population average), variance and correlation (between measurements on the same subject), e.g.

$$\begin{aligned} \mathrm{E}(Y_{ij}) = \mu_{ij} &= \beta_0 + \beta_1 t_{ij} \\ \mathrm{Var}(Y_{ij}) &= \sigma^2 \\ \mathrm{Corr}(Y_{ij}, Y_{ik}) &= \rho(\mu_{ij}, \mu_{ik}; \boldsymbol{\alpha}) \end{aligned}$$

Mixed Model: Models individual curves, e.g.

$$Y_{ij} = (\beta_0^* + b_{i0}) + (\beta_1^* + b_{i1}) t_{ij} + \varepsilon_{ij}$$

$$(b_{i0}, b_{1i})' \overset{iid}{\sim} \left(\mathbf{0}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right) \text{ ind. of } \varepsilon_{ij} \overset{iid}{\sim} (0, \sigma^2)$$

Transition Model: Models the present in terms of the past, e.g. $(q = 1)$

$$\begin{aligned} Y_{ij} &= \beta_0^{**} + \beta_1^{**} t_{ij} + \epsilon_{ij} \\ \epsilon_{ij} &= \alpha \epsilon_{ij-1} + \xi_{ij}, \;\; \xi_{ij} \overset{iid}{\sim} (0, \tau^2), \epsilon_{i1} \sim (0, \sigma^2) \end{aligned}$$

The Three Approaches for the Linear Model

Note that in the linear model, the β parameters in all three approaches have a marginal interpretation, i.e. in all three models do they measure the unit increase in the mean of Y for a unit increase in the corresponding covariate. For example:

Marginal Model:

$$\mathrm{E}(Y_{ij}) = \beta_0 + \beta_1 t_{ij}$$

Mixed Model:

$$\mathrm{E}(Y_{ij}) = \beta_0^* + \beta_1^* t_{ij}$$

Transition Model:

$$\mathrm{E}(Y_{ij}) = \beta_0^{**} + \beta_1^{**} t_{ij}$$

The Three Approaches and the Generalized Case

Consider now a generalized linear model, say a logistic model for the probability of having an infection (Y), given (no) vitamin A deficiency (x). Cross-sectional model:

$$\text{logit}\,\text{Pr}(Y_{ij}=1)=\beta_0+\beta_1 x_{ij}.$$

Marginal Model: Models mean, variance and correlation

$$\begin{aligned}\text{logit}\,\text{Pr}(Y_{ij}=1)=\text{logit}\,\mu_{ij} &= \beta_0+\beta_1 x_{ij}\\ \text{Var}(Y_{ij}) &= \mu_{ij}(1-\mu_{ij})\\ \text{Corr}(Y_{ij},Y_{ik}) &= \rho(\mu_{ij},\mu_{ik};\boldsymbol{\alpha})\end{aligned}$$

Mixed Model: Each individual has its own propensity for an infection

$$\begin{aligned}\text{logit}\,\text{Pr}(Y_{ij}=1|b_i)=(\beta_0^*+b_i)+\beta_1^* x_{ij}\\ b_i \overset{iid}{\sim} (0,\sigma_1^2)\end{aligned}$$

Transition Model: The probability for an infection depends on whether there was an infection at the last visit

$$\text{logit}\,\text{Pr}(Y_{ij}=1|Y_{ij-1},\ldots,Y_{i1}) = \beta_0^{**}+\beta_1^{**}x_{ij}+\alpha Y_{ij-1}$$

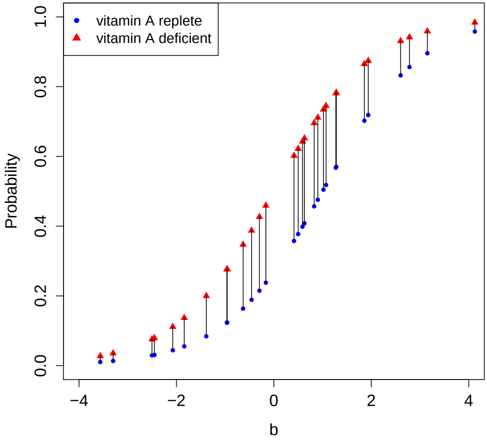
The Three Approaches and the Generalized Case

Now, the β parameters have quite different interpretations and will typically differ.

- β_1 is the log-odds ratio of infection between vitamin A deficient and replete children. It is a *population-averaged parameter* (marginal interpretation).
- β_1^* is the log-odds ratio of infection when a child is deficient relative to when that same child is not (conditional interpretation, conditional on individual propensity to infection). The resulting change in absolute risk depends on the baseline rate for that child.
- β_1^{**} is the log-odds ratio of infection for vitamin A deficiency versus repletion among the group of children free of infection (resp. infected) at the last visit (conditional interpretation, conditional on infection status at last visit).

For the logistic regression: (Neuhaus et al, 1991; Zeger et al, 1988)
$|\beta_1| \leq |\beta_1^*|$ with equality iff $\beta_1^* = 0$, and an increase in discrepancy with σ_1^2.
If $b_i \overset{iid}{\sim} N(0, \sigma_1^2)$, then $\beta_1 \approx (c^2\sigma_1^2 + 1)^{-1/2}\beta_1^*$ with $c^2 \approx 0.346$.
(Analogous component-wise for a vector $\boldsymbol{\beta}$.)

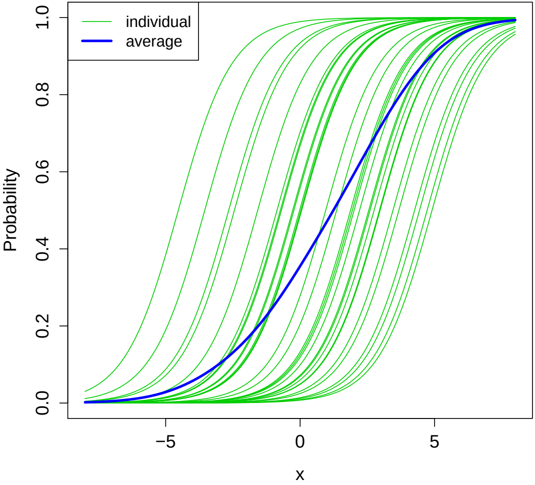
Marginal and Mixed Logistic Model



Marginal probability: $\rightarrow$ Marginal odds-ratio $\neq \exp(\beta_1^*)$

$$\Pr(Y_{ij}=1)=\int \Pr(Y_{ij}=1|b_i)dF(b_i)=\int \frac{\exp(\beta_0^*+b_i+\beta_1^* x_{ij})}{1+\exp(\beta_0^*+b_i+\beta_1^* x_{ij})} f(b_i)db_i$$

Marginal and Mixed Logistic Model



- Steep rate of increase in individual curves attenuated in average curve.
- Phenomenon well-known in related errors-in-variables regression problem.

The Three Approaches - Some Pros & Cons

Marginal Models:

- + Separate modeling of mean and correlation. Correlation model does not change interpretation of β parameters.
- ± Are appropriate for inference about the population mean.
- + Only requires specification of the first two moments, not the entire likelihood (less assumptions).
- – No likelihood-based inference (but instead: GEE).
- + Can easily accommodate unequally spaced time points and unbalanced data.

Mixed Models:

- + Can make inference about individuals rather than about population averages.
- + Can easily accommodate unequally spaced time points and unbalanced data.
- + Flexible, can easily incorporate additionally clustered data, smooth functions (penalized spline smoothing) etc.
- ± Parsimonious modeling of covariance. But random effects imply specific correlation structure, less flexibility.
- + Allows likelihood-based inference.
- – Fitting of models often hard (especially for generalized case).

The Three Approaches - Some Pros & Cons

Transition Model:

– Most meaningful for equally spaced time points t_{ij}. What about unequally spaced data? Missing data?

+ Might be a very meaningful way to think of the underlying process in some cases (e.g. for categorical data when thinking of transition between "states")

+ Allows likelihood-based inference (typically conditional on first q observations).

Missing Data

Missing data patterns:

- Dropout / loss-to-follow-up: Whenever Y_{ij} is missing, so are Y_{ik} for all $k \geq j$.
- Intermittent missing values

Missing data mechanisms: (Rubin, Biometrika, 1976)
Let $\mathbf{Y}$ indicate the complete data, with $\mathbf{Y}_{obs}$ the observed part and $\mathbf{Y}_{mis}$ the missing part of $\mathbf{Y}$. Define the missing data indicator $R_{ij} = 1$ if Y_{ij} is missing, and $= 0$ otherwise. (Note: sometimes $\mathbf{R}$ is defined the other way around.)

- Missing completely at random (MCAR): $p(\mathbf{R} \mid \mathbf{Y}) = p(\mathbf{R})$, so that the observed data are a completely random sample of the complete data
- Missing at random (MAR): $p(\mathbf{R} \mid \mathbf{Y}) = p(\mathbf{R} \mid \mathbf{Y}_{obs})$, so that the missing data mechanism does not depend on the actual missing values
- Not missing at random (NMAR): $p(\mathbf{R} \mid \mathbf{Y})$ depends on $\mathbf{Y}_{mis}$, so that whether or not an observation is observed depends on the quantities that you were not able to observe. (Also called informative missingness.)

Keep in mind for now. More later.

Longitudinal and Other Data

- **Multivariate Data**: (Balanced) Longitudinal Data can be viewed as a type of multivariate data. But with a special correlation structure!
- **Time Series**: Also correlation over time, but typically one (or a few related) longer time series. Longitudinal Data simpler, as there are (typically) independent subjects → borrow strength over subjects, more robust to model assumptions about correlation structure.
- **Hierarchical / Multi-level Data**: Similar nested structure and similar approaches (random effects etc.), but without the temporal structure.
- **Spatial Data**: 2-D / 3-D, no inherent ordering, not necessarily repeated measurements. But many similar approaches to modeling correlation: Marginal models, Gaussian random effects / fields, Markov chains / random fields on grids
- **Functional Data**: There are approaches to model longitudinal data as functional data (in time). → John will talk about that.

Sometimes we can learn from different areas . . . .