BIOS6643 Longitudinal

L1 Introduction

EJC

Department of Biostatistics & Informatics

L1 Introduction

Time series and longitudinal data

Time series data types and examples

Longitudinal data types and examples

Simple clustered/longitudinal analyses

Usual assumptions for longitudinal models

Longitudinal designs and power - an initial glimpse

Learning Objectives

1. Understand why we need special methods
2. Discuss example datasets that are longitudinal
3. Discuss time series vs. longitudinal; formats for longitudinal data
4. Understand the assumptions for longitudinal models
5. Review analyses of longitudinal data with two time points

Questions

- What makes longitudinal data different, so that we need special methods?
- What are clustered data?
- What are benefits of longitudinal models? (Or models for clustered data)
- Why are longitudinal methods not used more?

Longitudinal designs

- Designed experiments and observational studies can be applied to cross-sectional or longitudinal settings. Here, they are defined for the latter.
- A controlled experiment involves an intervention, while an observational study does not.
- In many cases a controlled experiment will have one or more true treatment groups, along with a 'control' group that either receives some type of placebo, or does not receive any treatment.
- **See the course notes for more detail on designed experiments versus observational studies.**

Time series and longitudinal data

Time series methods (generally)...

- focus on modeling one process over time (i.e., one observation taken at each time point, across time).
- focus on predicting values of future occurrences.

Generally, time series data can be found everywhere, including: stock prices, temperature, birth and mortality rates, health data for individuals (e.g., blood pressure), just to name a few areas.

Longitudinal methods (generally)...

- Involve measurements on multiple subjects.
- Assume that the correlation structure is the same across subjects but that responses are independent between subjects.

Often fewer time points for longitudinal data than time series data. Although analytic methods for time series and longitudinal data differ, they do have common elements, and the underlying processes that generate the data are often similar.

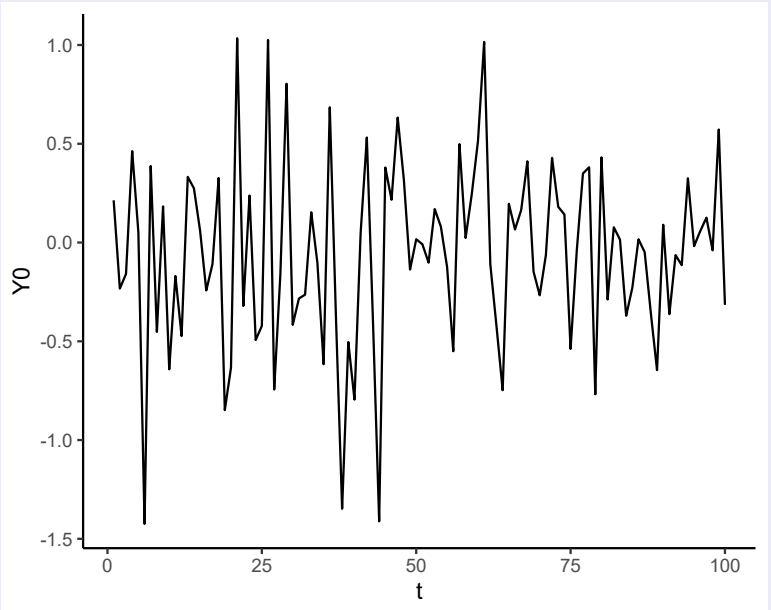
Time series data types and examples

Stationary processes

- A stationary process $\{Y_t\}$ has a constant mean (expected value) and finite 2nd moment for all times t, and the correlation between Y_t and Y_{t+h} does not depend on t, for all h.
- Below, data for stationary processes were simulated for the model, $Y_t = \mu + \epsilon_t$ where μ is the mean and ϵ_t are errors that are identically but not necessarily independently distributed.

Example 1: Stationary process (iid error)

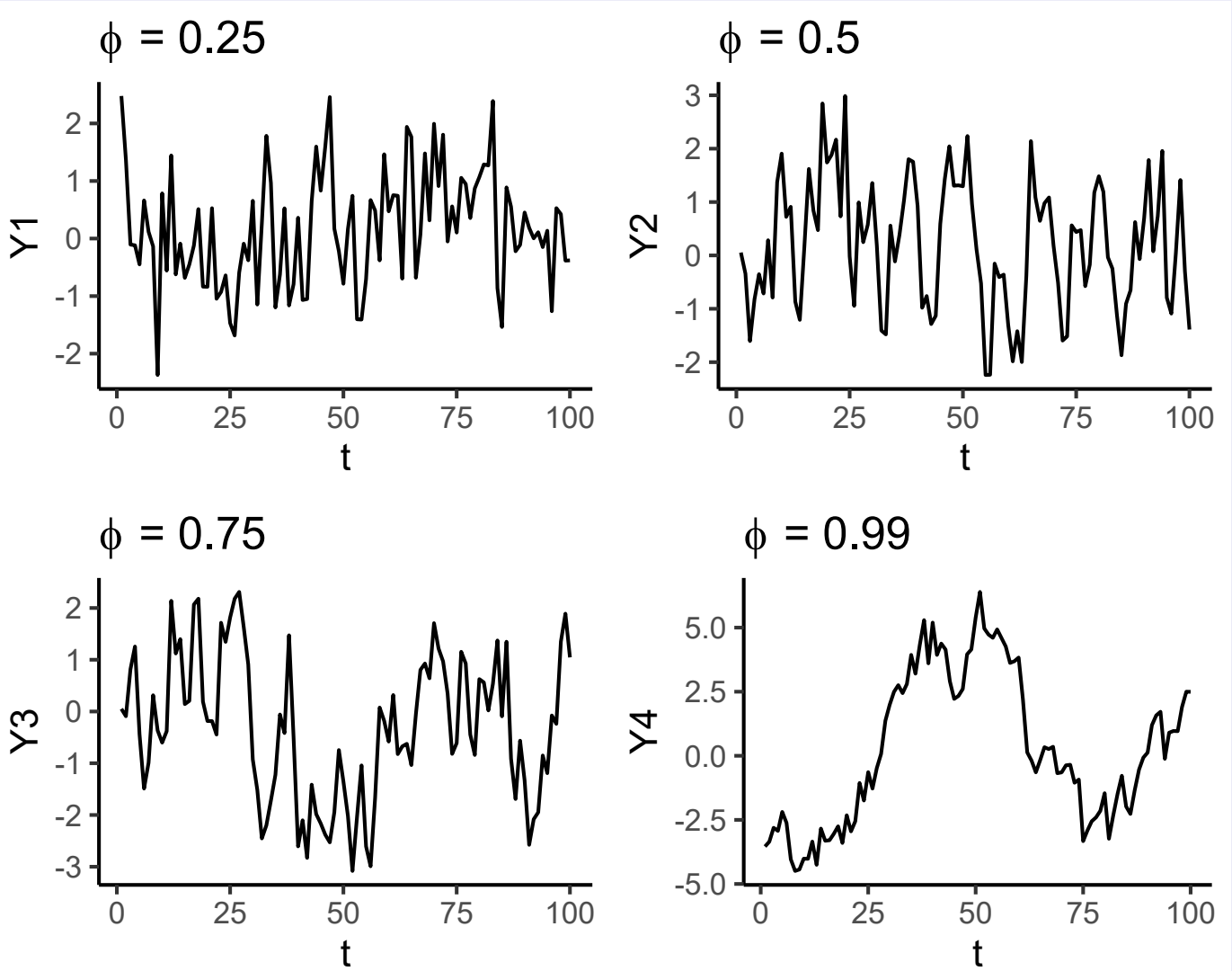
For the simulated data, $\mu = 0$ and $\epsilon_t \sim \mathcal{N}(0,\ 0.46)$ for all t.



Example 2: Stationary process (correlated error)

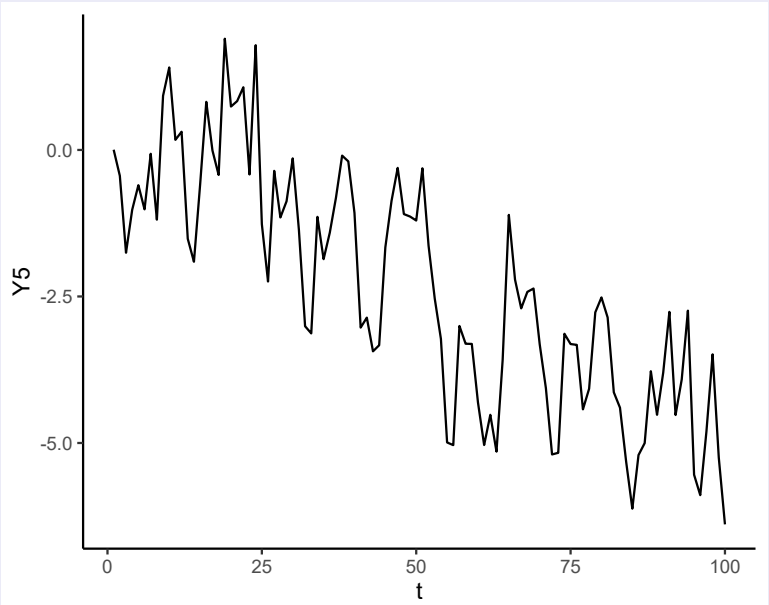
- Data below were generated using $\mu = 0$ and errors that followed a first-order autoregressive $(AR(1))$ process: $\epsilon_t = \phi\epsilon_{t-1} + Z_t$ and Specifically, $Z_t \overset{iid}{\sim} \mathcal{N}(0,\ 0.46)$, for all t.
- **Notes on AR(1) processes**:
  1. Errors ϵ_t are identically distributed but not independent
  2. Must have $|\phi| < 1$ for stationary process
  3. The higher the value of $|\phi|$, the higher degree of correlation between responses from day to day

Example 2: Stationary process (correlated error)



Example 3: Processes with trend and correlated errors

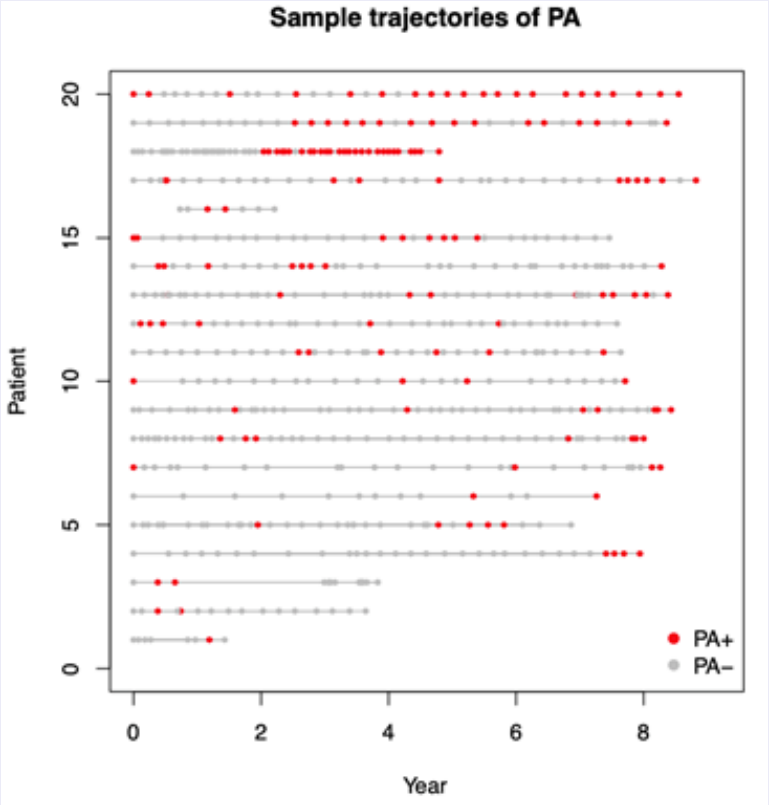
- $AR(1)$ process with linear time trend.
- $Y_t = \beta_0 + \beta_1 t + \epsilon_t$, $\beta_0 = 0$, $\beta_1 = -0.05$, $\epsilon_t \sim AR(1)$ (as in **Example 2**, last page, with $\phi = 0.25$)



Random walks - see course notes (includes Example 5)

Example 4: Retrospective observational studies

- Longitudinal trajectories of pseudomonas (PA) in EPIC study
  - 1734 children enrolled in the EPIC observational study who were pseudomonas negative (PA-) for early pseudomonas infection control observational study (EPIC)
  - Median followup time is 7.8 years ($Q1 - Q3: \ 6.3 - 8.3$)
  - One of the questions of interest was finding factors associated with progression of PA; A secondary outcome of interest: time to first pulmonary exacerbation in EPIC trial



Example 5: Prospective observational studies

STEPPED-CARE randomized trial.

- A behavioral intervention was tested versus usual care in patients with lung or head and neck cancer.

need some figures

Example 6: Chronic fatigue syndrome study

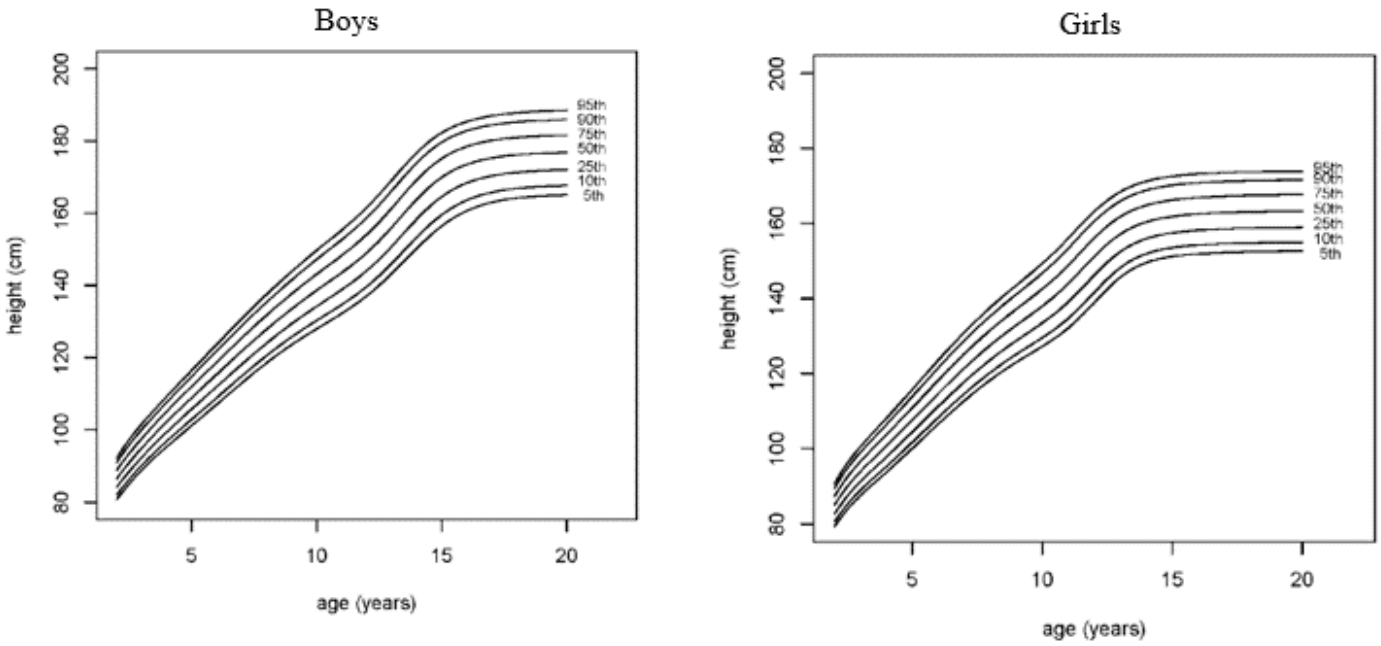
Complement levels and Chronic fatigue syndrome (CFS) data (Sorensen et al., 2003).

- Involved measuring complement split products (biological markers) over time.
- In this case, groups involved those with or without CFS, and thus repeated measures only involved time.
- A special covariance structure was used to model the repeated measures since measurement times were unequally spaced.

Example 7: Growth curve data

- Graphs for height as a function of age for boys and girls aged 2 to 20 years
- Constructed in R after obtaining growth data from the CDC. For more information, please see growcharts.
- These data show that girls approach their maximum height much more quickly than boys. The y-axis scales were made the same for easier comparison between graphs.
- Each curve is a percentile estimate as a function of age. We could create confidence bands for each percentile curve.
- If the curves are estimated using a lot of data, the widths of the bands should be narrow. Doctors look for dramatic changes between visits.
- The curves here may not be representative of all populations (e.g., differences due to race).

Example 7: Growth curve data



See course notes for more data examples

Formats for longitudinal data

This section covers

- how to set up data for "univariate" versus "multivariate" analysis
- types of variables and notation for variables in longitudinal data versus "factorial" data
- **See course notes for more information**

Example 8, 9, & 10: Cluster data

Example 8: After an exercise challenge performed on 20 subjects, resting heart rates are monitored at 5 minute intervals for one hour. How are data clustered?

Example 9: Families are selected to participate in a survey regarding health insurance.
Each member of the family will be included in the study.

Example 10: arm length and leg length growth are measured for subjects once a year for 10 years, and then modeled with a linear mixed model.

What we've already done!

- Experiments with pre-post measurements have 2 measurements on each subject over time. When there are only 2 measurements, the analysis simplifies when the difference is considered, as the analysis is reduced to one measurement per subject. Simple methods can then be used (e.g., paired t-test).
- Longitudinal models can still be beneficial here! But we'll discuss that later. For now, we consider simplified models.
- Let's take a closer look at the underlying models when we use a difference score or take the baseline-as-covariate approach.

Change-score model and Baseline-as-covariate model

Change-score model

$$Y_{i1} = Score_{pre};\ Y_{i2} = Score_{post}$$
$$\Delta_i = Y_{i1} - Y_{i2} = \beta_0 + \beta_1 x_i + \epsilon_i$$

Baseline-as-covariance model

$$Y_{i2} = \beta_0 + \beta_1 Y_{i1} + \beta_2 x_i + \epsilon_i$$

We allow the slope of the baseline value to be anything (based on fit).

Example for discussion: cholesterol data

Any other type of simple clustering, with 2 responses per cluster can be analyzed similarly. (E.g., pairing by married couple, pairing by year of measurement.)

- Assumption 1: Responses between subjects are independent.
  - If there are clear violations to the assumption, and data are available, then a random term could be added to deal with this non-independence.
  - For example, if a class is used for the sample, and there are several pairs of siblings in the class, a random term identifying family could be added to the model. (Lack of fit and lack of independence are related!)
- Assumption 2: There is a common covariance structure between all subjects, and the covariance parameters have the same value between subjects.
  - This assumption is usually not tested. However, to properly estimate covariance parameters, several subjects are needed (just as data for several subjects are needed to estimate a common population mean).
  - In some cases, homogeneous groups within the study may be identified (but heterogeneous between groups). With sufficient group sample sizes, group-specific covariance parameters can be put in the model and estimated.

Longitudinal designs and power - an initial glimpse

- Consider an experiment designed to compare two treatments. Two common approaches:
  - A1: Use independent samples (randomly assign some subjects one treatment, and some the other). For A1, we often use a 2-independent sample t-test
  - A2: Have all subjects have one treatment and then have them all take the other (e.g., use a crossover design to eliminate confounding effects related to time). For A2, a paired t-test.
- A study/experiment involving changes within subjects (e.g., analyzed with a paired t-test) is often more powerful than a study using independent samples.

- The general formula for the variance for the difference in means suggests why this may be expected (when correlations between responses within subjects are positive):

$$Var[\bar{Y}_1 - \bar{Y}_2] = Var[\bar{Y}_1] + Var[\bar{Y}_2] - 2Cov[\bar{Y}_1,\ \bar{Y}_2]$$

- Often there are many factors not of interest that distinguish the two independent samples, while for the paired data, the difference in responses is due more to the treatment alone and not to other factors, since we're using the same subjects.
- The same principle generalizing to multiple times and longitudinal data in general (e.g., air pollution study); subject serve as their own controls.
- But paired/longitudinal designs may not always be better. In some cases a short cross-sectional study/experiment involving many subjects may be more feasible and cost-effective.

Summary

1. Why do we need special methods?
2. Discussed example datasets that are longitudinal
3. Discuss time series vs. longitudinal; formats for longitudinal data
4. Assumptions for longitudinal models
5. Analyses of longitudinal data with two time points