1 Some questions

- What makes longitudinal data different, so that we need special methods to analyze it?

- Why are longitudinal methods not used more?

- What are clustered data?

2 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.

2.1 Designed experiments

- Clinical trials and basic science experiments are a few types of designed experiments.

- Common design types

  - Parallel: subject are randomly assigned a treatment and continue with it for the duration of the experiment.

  - Crossover: subjects are randomly assigned treatments; at some point during the experiment, subjects then switch treatments (at least once).
    - Size (e.g., 2-period, 2-treatment; 3-period, 3-treatment)
    - Washout period(s)
    - Randomization (to eliminate confounding effects associated with time).
    - Parallel or Crossover best? Depends…

2.2 Observational studies

- Retrospective study.

- Prospective study.

- Epidemiologic time-series study.

3 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 analytical methods for time series and longitudinal data differ, they do have common elements, and the underlying processes that generate the data are often similar.

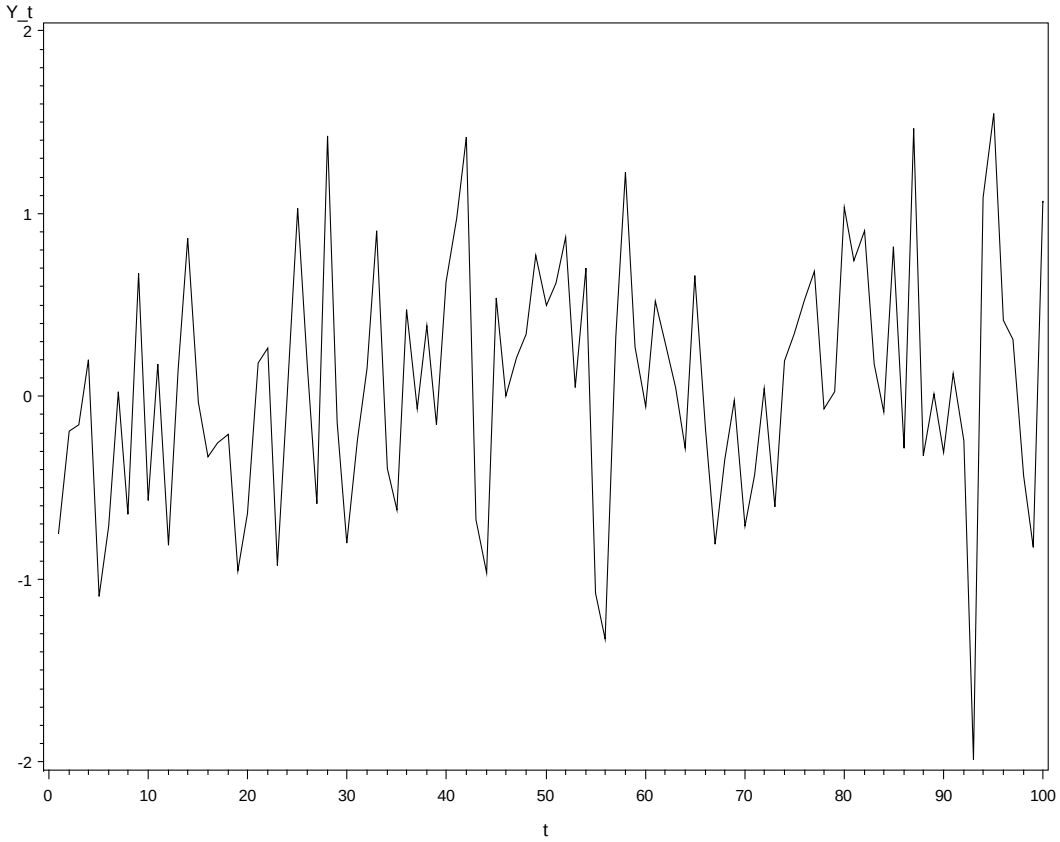
3.1 Time series data types and examples

3.1.1 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 using SAS and the model $Y_t = \mu + \varepsilon_t$, where μ is the mean and ε_t are errors that are identically but not necessarily independently distributed.

<u>Example 1</u>: Stationary process with identically and independently (iid) distributed errors. For the simulated data to the right, μ=0 and ε_t ~ Normal with mean 0 and variance 0.46 for all *t*.



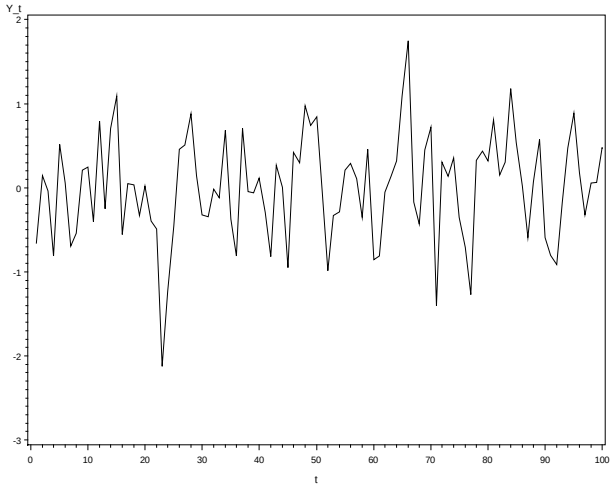
<u>Example 2</u>: Stationary process with correlated errors.

- Data below were generated using $\mu=0$ and errors that followed a first-order autoregressive [AR(1)] process: $\varepsilon_t = \phi\,\varepsilon_{t-1} + Z_t$ and $Z_t \sim iid$ for all t. (Specifically, Z_t ~Normal with mean 0 and variance 0.46.)

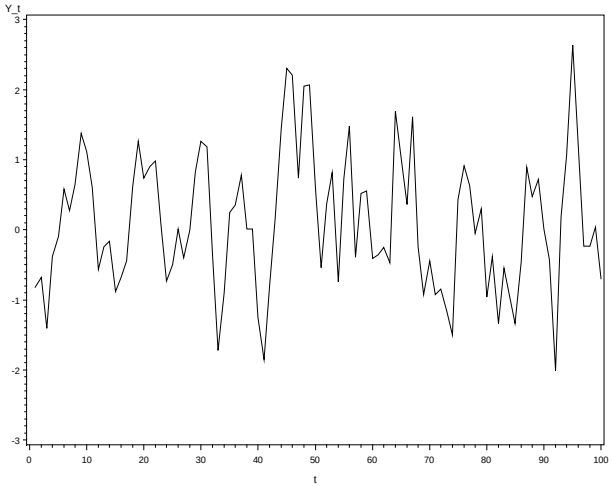
- A few notes on AR(1) processes:

  (i) Errors ε_t are identically distributed but not independent
  (ii) Must have $|\phi|<1$ for stationarity
  (iii) The higher the value of $|\phi|$, the higher degree of correlation between responses from day to day

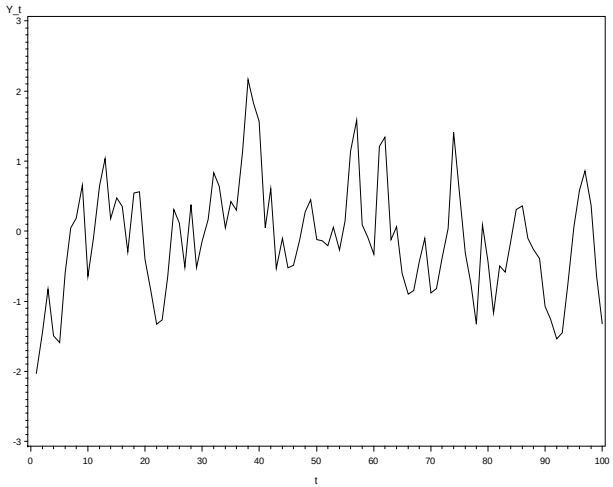
ϕ=0.25



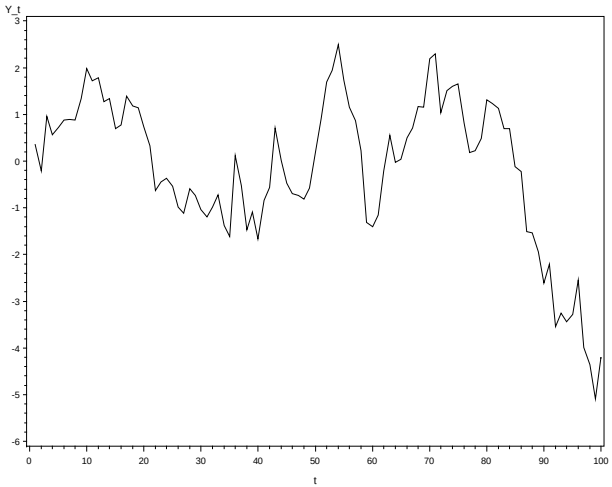
ϕ =0.5



ϕ =0.75

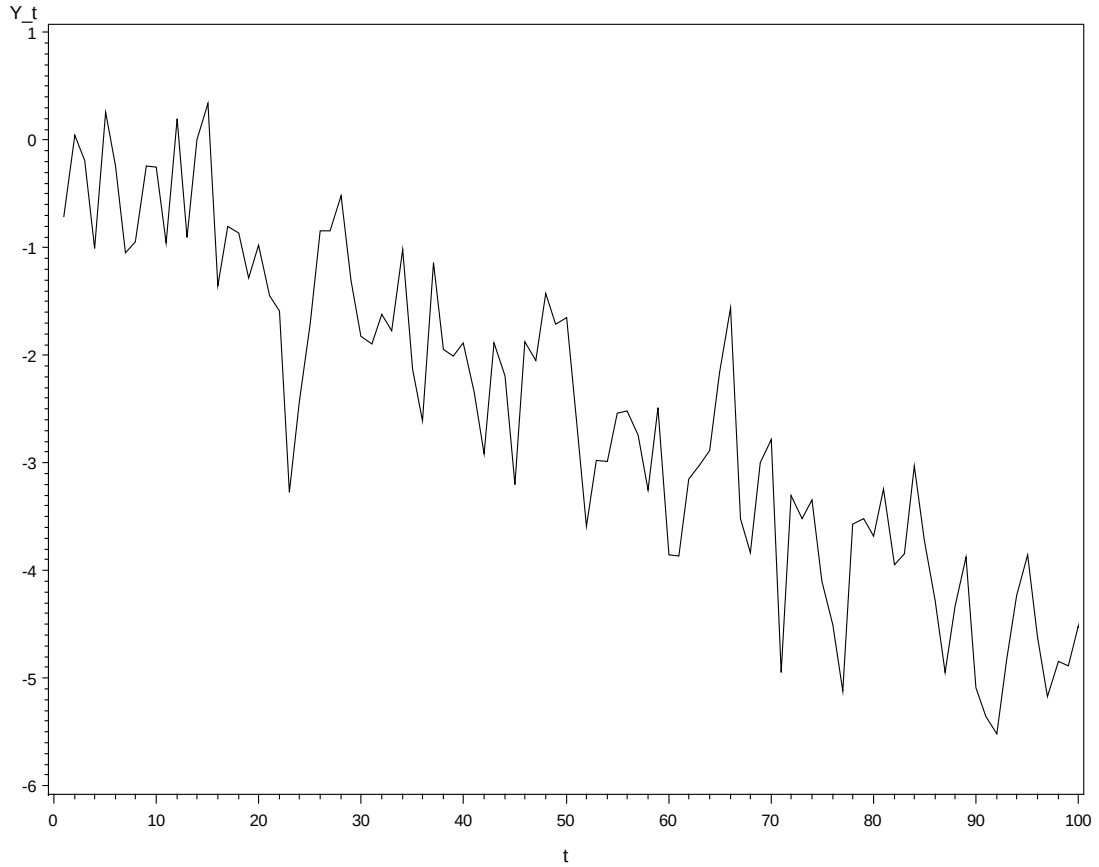


ϕ =0.99

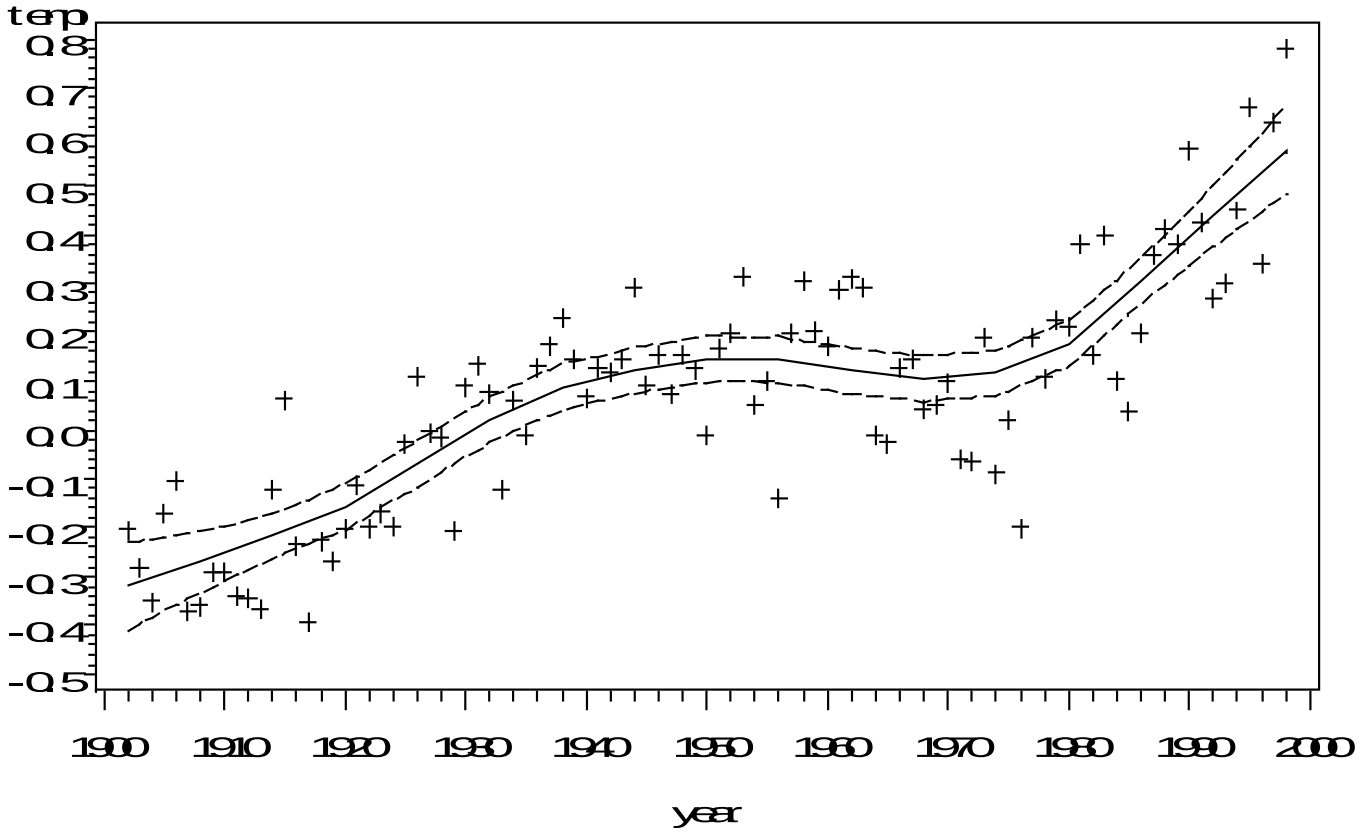


3.1.2 Processes with trend and correlated errors

- *Example 3*: AR(1) process with linear time trend. $Y_t = \beta_0 + \beta_1 t + \varepsilon_t$, β_0=0, β_1= –0.05, ε_t ~AR(1) (as in Ex. 2, last page, with ϕ =0.25)



- *Example 4*: Global temperature data, 20th century, with nonparametric regression fit (95% pointwise confidence bands for the mean in dashed lines).



3.1.3 Random walks

- A random walk is a process that involves movement in random directions.

- For example, the path traced by a molecule as it travels in a liquid or a gas, the search path of a foraging animal and the price of a fluctuating stock can all be modeled as random walks (Wikipedia).

- Example 5: the following graph depicts several realizations of the following random walk model. This is equivalent to the following: flip a coin; if heads, go forward and to the left; if tails, goes forward and to the right; flip coin again and use the same decision rule; keep repeating.

Random walk example:

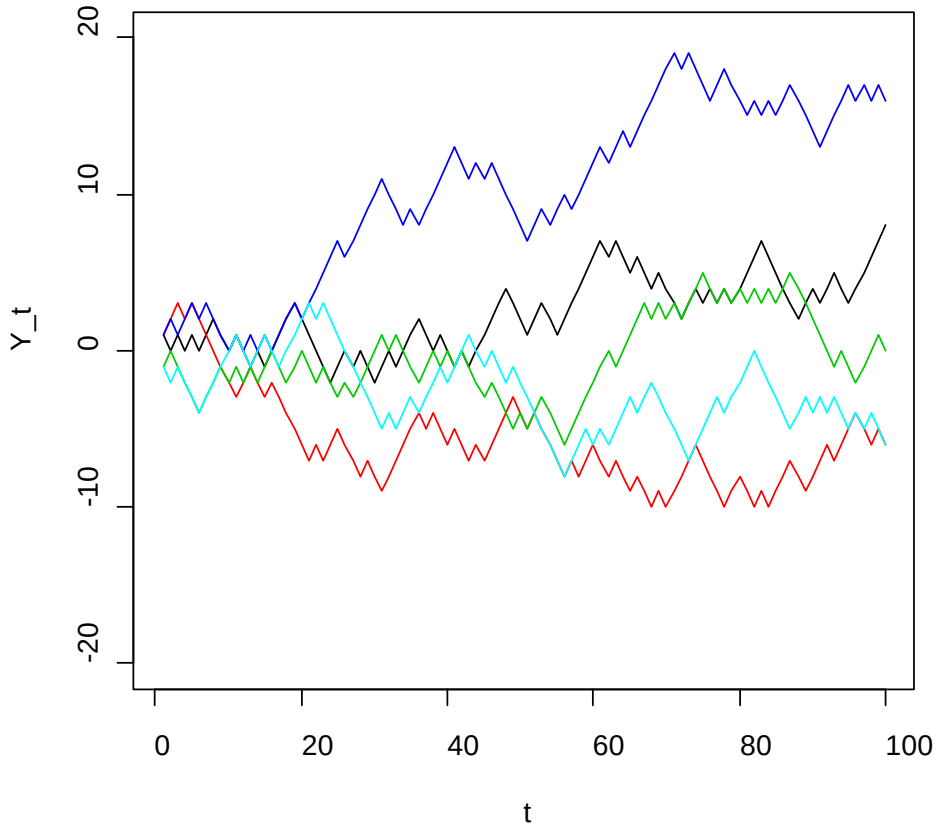
$$Y_t = Y_{t-1} + B_t = \sum_{i=1}^{t} B_i$$

B_t = 1 with probability p
= –1 with probability 1–p

Here, $Y_0 = 0$, $p = ½$.

This process is an example of a *Markov chain*, since

$P(Y_{t+1} = y \mid Y_0, Y_1, \ldots, Y_t)$
$= P(Y_{t+1} = y \mid Y_t)$



3.2 Longitudinal data types and examples

The following examples show some of the common categories of longitudinal data, and most of these are real examples from my job.

3.2.1 Retrospective observational studies

Example 6: *Occupational Medicine – natural history of Beryllium disease.*

- Involved looking at workers in production plants that use(d) Beryllium metal to determine the progression of health for subjects that contract Chronic Beryllium Disease (CBD) versus those that only become sensitized to Beryllium (BeS).
- Several outcomes were measured in four main categories: blood and broncoalveolar lavage, pulmonary function (e.g., FEV_1, FVC), exercise physiology (e.g., $AADO_2$ Rest, Max VO_2). The main aim was to summarize the progression of illness for CBD subject versus BeS subjects, not due to aging.
- Complex data:
  - Different tests were done on different days.
  - Each subject had different number of tests done over time.
  - Records spanned several decades.

Example 7: MDR-TB study.

- Involved obtaining records from subjects who were initially treated at National Jewish Health for multiple-drug resistant tuberculosis.
- Difficult retrospective study: involved getting more information than just what was available in medical records at the hospital. Subjects or their doctors were contacted to determine their health status up to many years after they had been discharged from the hospital. In some cases, subjects had passed away or it was just not possible to contact them.
- Analyses
  - Logistic regression for treatment success (evaluated based on sputum tests taken during or soon after discharge from the hospital)
  - Survival analysis for survival from TB (using longer-term data from medical records and follow up).
- Although these are longitudinal data, the logistic regression did not use repeated measures, and survival analysis is typically considered in a category separate from longitudinal data analysis.

3.2.2 Prospective observational studies

A prospective study involves collecting information over time based on a pre-planned design, without an intervention. In a prospective longitudinal study, more planning is possible with respect to when measurements are taken.

<u>Example 8</u>: Kunsberg/Air pollution study.

- Involved students attending the K-8 school at National Jewish Health.

- Health and behavioral variables were collected on subjects over time, some daily (e.g., daily albuterol use) and some more intermittent (e.g., personal exposure estimate and biomarkers from urine samples).

- Concurrently, air pollution measures from fixed monitors and personal monitors were taken. The relationship between health and environmental were examined. See Rabinovitch et al. (2004, JACI; 2006 and 2011, AJRCCM).

3.2.3 Epidemiologic time series studies

This involves modeling of data at a larger, more aggregated level, typically with many days of observation. Data may fall more into the 'time series' class, although often a standard longitudinal model may still be used to fit the data.

<u>Example 9</u>: Relationship between hospital admission counts and PM_{10} in the San Luis Valley over 11 years.

- Involved determining the association between hospital counts at a medical center in Alamosa, Colorado that serves the greater San Luis Valley area, and concurrent PM_{10} concentrations – i.e., airborne coarse particulate matter.

- Although rural, airborne sand and dust particles in the valley can contribute to higher PM_{10} concentrations. (Don't forget, the Sand Dunes are down there!)

- To determine a more 'pure' relationship between health and PM_{10}, the model also accounted for temporal trends as well as other environmental factors such as meteorology. Both seasonal and long-term time trends were accounted for flexibly in the model by including spline terms.

3.2.4 Clinical trials

- A clinical trial is typically a controlled experiment involving human subjects, with the aim of determining whether a new drug or therapy is better than a standard of care, existing medication or placebo.

- Often subjects are randomized to a treatment group (which could include a control group), and then observed over time, or to a treatment sequence, in which case they receive multiple sequences, separated by washout periods.

- This will be discussed a bit more in Section 2. At my current job I have analyzed data from several clinical trials.

Example 10: One trial involved giving aspirin-allergic subjects an aspirin challenge, with eNO measurements coming just before the challenge, 1 day post, and 6 months post.

Example 11: Another FDA-funded trial involved a crossover design in which a dose-response curve was estimated based on subjects that took multiple doses of a drug designed to reduce eNO.

3.2.5 Basic science experiments

- Help give a better understanding of what drives or modifies certain diseases at the cellular level.

- Often, blood or biopsy samples are taken from humans or animals in order to carry out the experiments. Sometimes cell cultures are extracted from a sample and placed into separate wells so that different treatments can be applied, after which cell counts or other measures are made to determine how the treatments affect cellular chemistry and activity. Measures may also be taken over time.

- Not hampered by issues that affect observational studies (easier to get data as planned).

- Experiments can often only be performed with a small number of experimental units.

Example 12: *Complement levels and Chronic fatigue syndrome 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 13: *Nuclear factor-Kappa B data*, based on article by Bai et al., to appear in *PLoS One*.

- Macrophage samples from a human subject were put into four separate cell cultures, each one incubated with one of 4 treatments (combinations of BAY – Y/N; TB: Y/N), then observed over time. Thus, each subject sample had 16 measurements, over time and treatment ('doubly repeated measures' – we will examine in more detail later). Macrophages originate from WBCs; they are "big eaters" of cellular debris and pathogens.

- This was repeated for a number of subjects. There were several outcomes measured, one being the amount of M. tuberculosis in the culture for the given condition.

- One of the major findings of the experiment was that BAY treatment (an inhibitor of NFκB) reduces M. tuberculosis in samples treated with TB, with greater relative differences occurring over time.

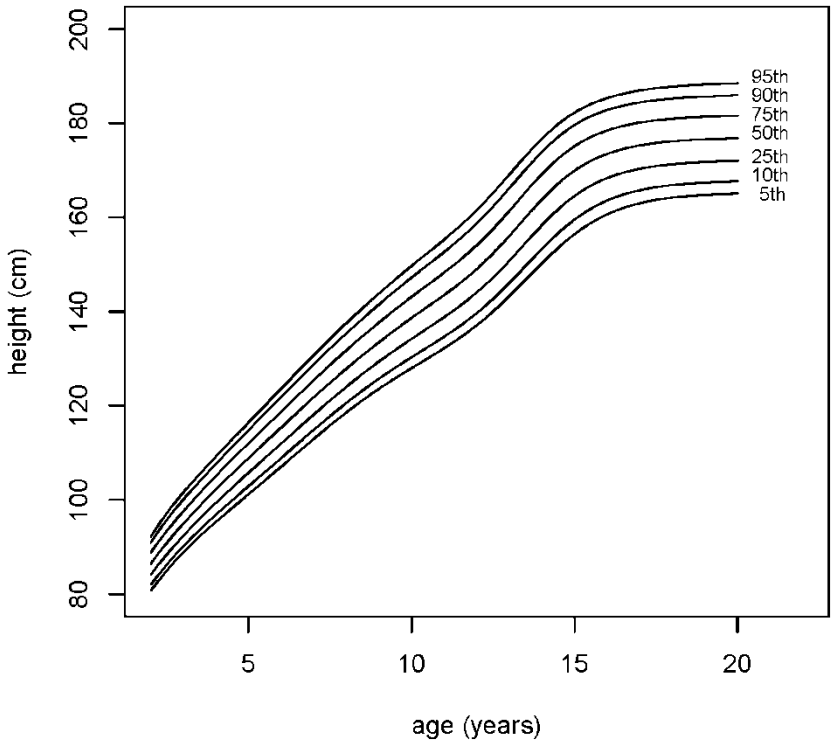
Example 14: *Myostatin data* (Taylor et al., 2001).

- The myostatin protein is an inhibitor of skeletal muscle mass.
- The experiment involved a 2×3 factorial treatment structure in a completely randomized design to determine effects of myostatin (Y/N) and Time (1, 2, 3 days) on protein levels in muscle cells.
- Muscle cells were taken from 24 mice and grown in separate tissue culture wells that each had a specific treatment (presence or absence of myostatin) and time of measurement.
- Protein degradation was observed over time, and samples treated with myostatin had greater protein degradation than those that did not. Data appear in Strand et al. (*Journal of Stat. Software*, 2004). Although this experiment involves time, it is not a longitudinal experiment in that it did not involve repeated measures.

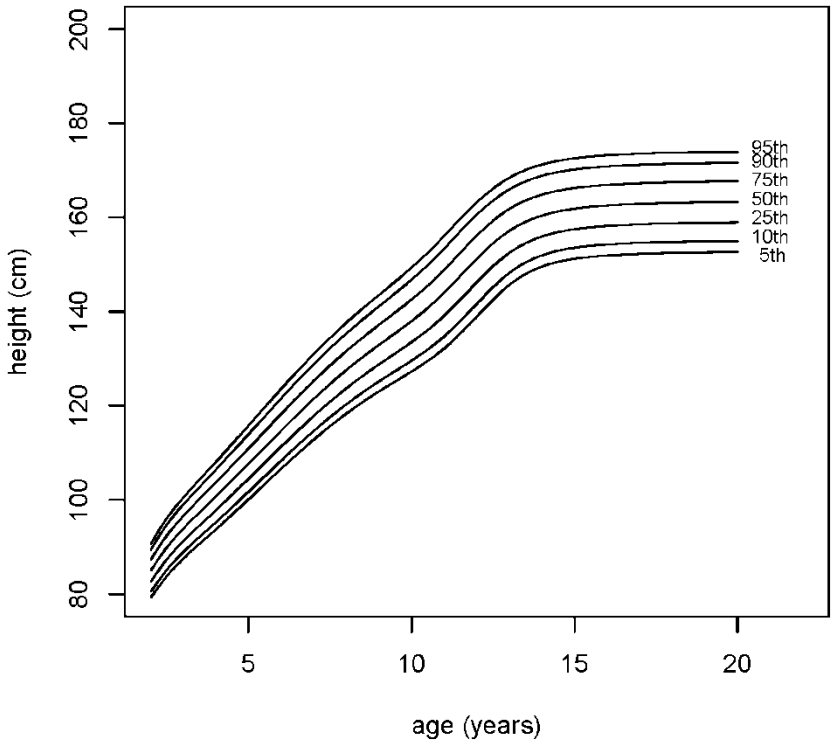
3.2.6 Growth curve data

Example 15: 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 (available at their website). For more information, please see http://www.cdc.gov/growthcharts/. 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).

Boys


height (cm)
80
100
120
140
160
180
200
5
10
15
20
age (years)
95th
90th
75th
50th
25th
10th
5th

Girls


height (cm)
80
100
120
140
160
180
200
5
10
15
20
age (years)
95th
90th
75th
50th
25th
10th
5th

4 Formats for longitudinal data

4.1 Dependent variable (or response or outcome variable)

'Multivariate' format (wide format; data are split)

Subject	Time 1	Time 2 …		Time r_i
1	Y_{11}	Y_{12}	…	Y_{1r_1}
2	Y_{21}	Y_{22}	…	Y_{2r_2}
…	...	…	…	…
n	Y_{n1}	Y_{n2}	…	Y_{nr_n}

'Univariate' format (long format; data stacked); Y_{ij} denotes the outcome; i=1,…, n denotes subject, j=1,…,r_i denotes time.

Subject	Time	Response
1	1	Y_{11}
1	2	Y_{12}
…	...	...
1	r_1	Y_{1r_1}
2	1	Y_{21}
2	2	Y_{22}
…	...	...
2	r_2	Y_{2r_2}
...	...	...
n	1	Y_{n1}
n	2	Y_{n2}
...	...	...
n	r_n	Y_{nr_n}

4.2 Independent variables (or predictors, or covariates)

- Covariates in a longitudinal study or experiment usually have 3 indices, at most – one to index subjects, one for time, and one for the type of variable (e.g., height, age, etc.). Types of covariates are discussed below.

  - Time-dependent:
    - x_{vij}
    - v denotes the variable (or predictor),
    - i denotes subject (i=1,…,n),
    - j denotes time.

- Time-invariant
  - E.g., race, gender
  - In some cases treatment is time invariant (e.g., for a parallel design).
  - If v is time-invariant, then $x_{vi1} = x_{vi2} = \ldots = x_{vir_i}$ for i=1,…, n.
- Subject-invariant
  - E.g., temperature on a given day for a study in the same location.
  - If variable j is subject-invariant, then $x_{v1j} = x_{v2j} = \ldots = x_{vnj}$ for times j=1,…, r.

- Categorical variables

  - Binary variables such as gender can be coded as a dummy or indicator variable (e.g., Female=1, Male=0, for gender).

  - A categorical variable with c levels can be uniquely coded with c–1 dummy variables.

  - E.g., x_1=0 for ‘L’, 1 for ‘M’, 0 for ‘H’; x_2=0 for ‘L’, 0 for ‘M’, 1 for ‘H’.

- Sometimes an index will either be redundant or not necessary, in which case it can be dropped. For example, for the time-invariant variables described above, we could drop j to yield x_{vi} .

4.3 Examples (dropping subject and time indices)

- *Example 16*: *growth study*.

  Y = height

  x_1=gender

  x_2=diet

- *Example 17*: *clinical trial, parallel design*.

  Y = eNO

  x_1= treatment (drug, control)

  x_2=gender

  x_3= baseline age

  x_4= time

  eNO = exhaled nitric oxide, a measure of airway inflammation; nitric oxide is a gaseous molecule produced by certain cell types in an inflammatory response.

4.4 Indices for variables and effects: longitudinal versus factorial models

- Responses for longitudinal data are often denoted as Y_{ij} (or Y_{it}), where i denotes subject and j (or t) denotes time. As long as each subject has a unique index across the study or experiment, these two indices are sufficient on a response, even if there is a class variable for groups of subjects that is on the right-hand side of the equation.

- For factorial data with two factors and replicates within each treatment combinations, we typically use something like Y_{ijk} to denote a response, where i denotes the level of the first factor, j denotes the level of the second factor, and k is the replicate. Since the replicate k refers to the specific treatment combination (i.e., it is not unique for subjects or objects across the study or experiment), it is important to keep this index on the response variable.

- Generally, statistical models can be written in different ways but you just have to make sure that the response variable and predictors correspond appropriately with respect to indices used.

5 Clustered data

Example 18: 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 19: Families are selected to participate in a survey regarding health insurance. Each member of the family will be included in the study.

Example 20: arm length and leg length growth are measured for subjects once a year for 10 years, and then modeled with a linear mixed model.

6 Simple clustered/longitudinal analyses (that 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). 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:

Y_{i1} = pre score
Y_{i2} = post score
$d_i = Y_{i2} - Y_{i1}$
$d_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

- Baseline-as-covariate model:

$$Y_{i2} = \beta_0 + \beta_1 Y_{i1} + \beta_2 x_i + \varepsilon_i \, .$$

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.)

7 Usual assumptions for longitudinal models

- 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.

8 Longitudinal designs and power – an initial glimpse

- Consider an experiment designed to compare two treatments. Two common approaches are to use independent samples (randomly assign some subjects one treatment, and some the other), or to have all subjects have one treatment and then have them all take the other (could be done using a crossover design to eliminate confounding effects related to time).

- For the first approach, we often use a 2-independent sample t-test, and for the second, 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):

$$Var\left(\bar{Y}_1 - \bar{Y}_2\right) = Var\left(\bar{Y}_1\right) + Var\left(\bar{Y}_2\right) - 2Cov\left(\bar{Y}_1, \bar{Y}_2\right).$$

- The reason why this is often the case is because 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 applies to longitudinal designs in general when the 'treatment' changes within subjects. For example, in the air pollution study (here, air pollution is the 'treatment'), subjects are exposed to different levels of air pollution day after day. We examine changes in their health and this provides a relatively powerful way to examine how air pollution is associated with health because the children serve as their own controls.

- This is not to say that paired/longitudinal designs are always better. In some cases a short cross-sectional study/experiment involving many subjects may be more feasible and cost-effective, particularly if the cost of getting an additional subject is more realistic or economical than keeping the same subject in the study for a longer period of time.