The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

1863-15

Advanced School and Conference on Statistics and Applied Probability in Life Sciences

24 September - 12 October, 2007

Longitudinal and Functional Data Analysis
An overview of Longitudinal Data Analysis

Anestis Antoniadis

LJK- Universite' Joseph Fourier
Grenoble, France

Longitudinal and Functional Data Analysis

An overview of Longitudinal Data Analysis

Anestis Antoniadis
LJK-Université Joseph Fourier

Trieste, October 2007

Outline

- Introduction
- Motivating examples
- Simplest models
- Estimation
- More advanced models
- Inference
- Computing

Introduction

More and more data are being collected over time on the same individual (i.e. subject, animal, sample). Both FDA and LDA are concerned with the analysis of such type of data.

Measurements treated in the FDA literature typically are recorded by high frequency automatic sensing equipment, whereas those treated in the LDA literature are more typically sparsely, and often irregularly, spaced measurements on human or other biological subjects.

The aims of the analysis are also often somewhat different:

- those of FDA tend to be exploratory to represent and display data in order to highlight interesting characteristics, perhaps as input for further analysis
- those of LDA have a stronger inferential component.

Common aims

Despite these differences in focus, there are many common aims, among them are the following:

- Characterization of average or "typical" time course.
- Estimation of individual curves from noisy and, often in LDA studies, sparse data. Functionals of these curves, such as derivatives and locations and values of extrema, are sometimes also of interest.
- Characterizing homogeneity and patterns of variability among curves, and identifying unusual ones.
- Assessing the relationships of shapes of curves to covariates.

As we shall see, many of these objectives entail smoothing individual curves, either explicitly or implicitly.

Longitudinal Studies

The defining characteristic of longitudinal studies is that subjects are measured repeatedly over time.

This is in contrast to cross-sectional studies where a single outcome is measured for each individual.

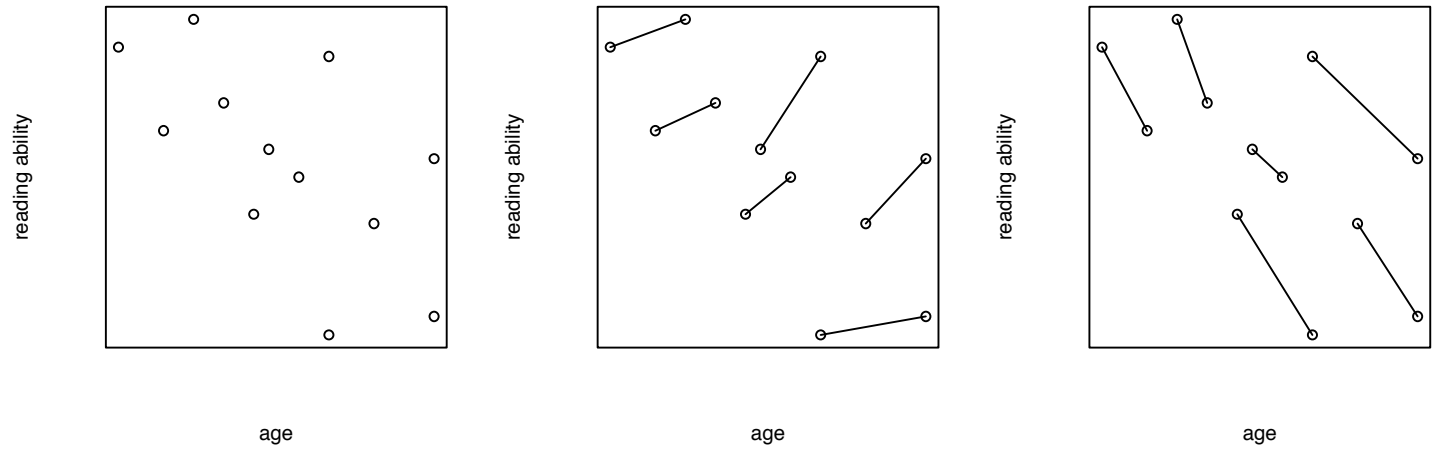
Data are collected over time because interest focuses on what happens over time! In a LDA one can investigate:

- changes over time within individuals (*age effects*)
- differences among subjects in their baseline levels (*cohort levels*)

Why collect longitudinal data?

To separate **age effects** from **cohort effects** .

Hypothetical data on reading ability against age:



- age effects : changes over time within subject
- cohort effects : differences between subjects at baseline

Examples of Longitudinal data

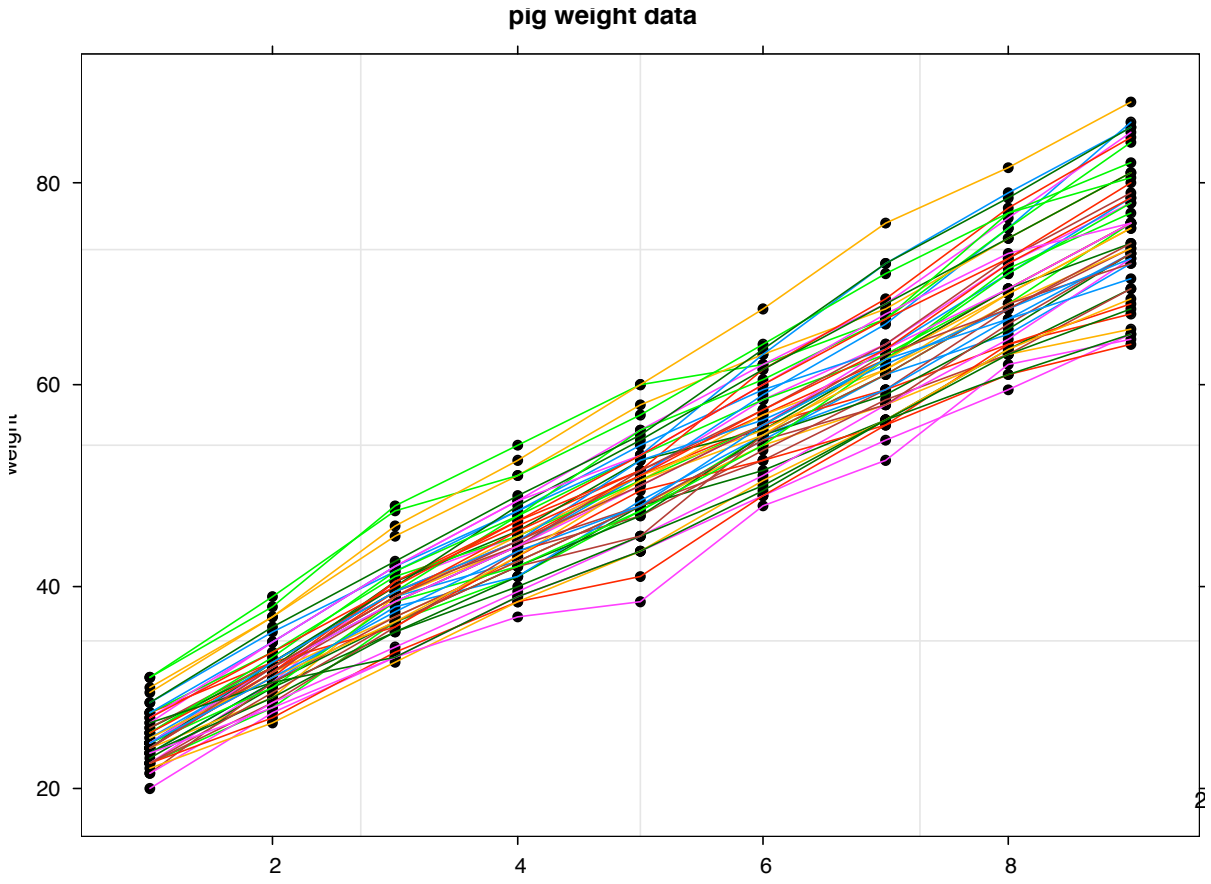
Pig weight data

- A cohort of $m = 48$ pigs was studied
- Weight measurements were obtained from each pig at $n_i = 9$ successive weeks

The scientific aims of the study is to characterize the growth patterns of these pigs.

A scatterplot of the weights against their corresponding week number with lines drawn connecting those measurements that belong to the same pig is as follows

Pig weights



"Orthodontic study data" (Potthoff and Roy (1964))

- involved 27 children, 16 boys and 11girls
- on each child, distance(mm) from the center of the pituitary to the pteryomaxillary fissure (two points that are easily identified on x-ray exposures of the side of the head) measured at ages 8, 10, 12, and 14 years of age
- A measure of growth

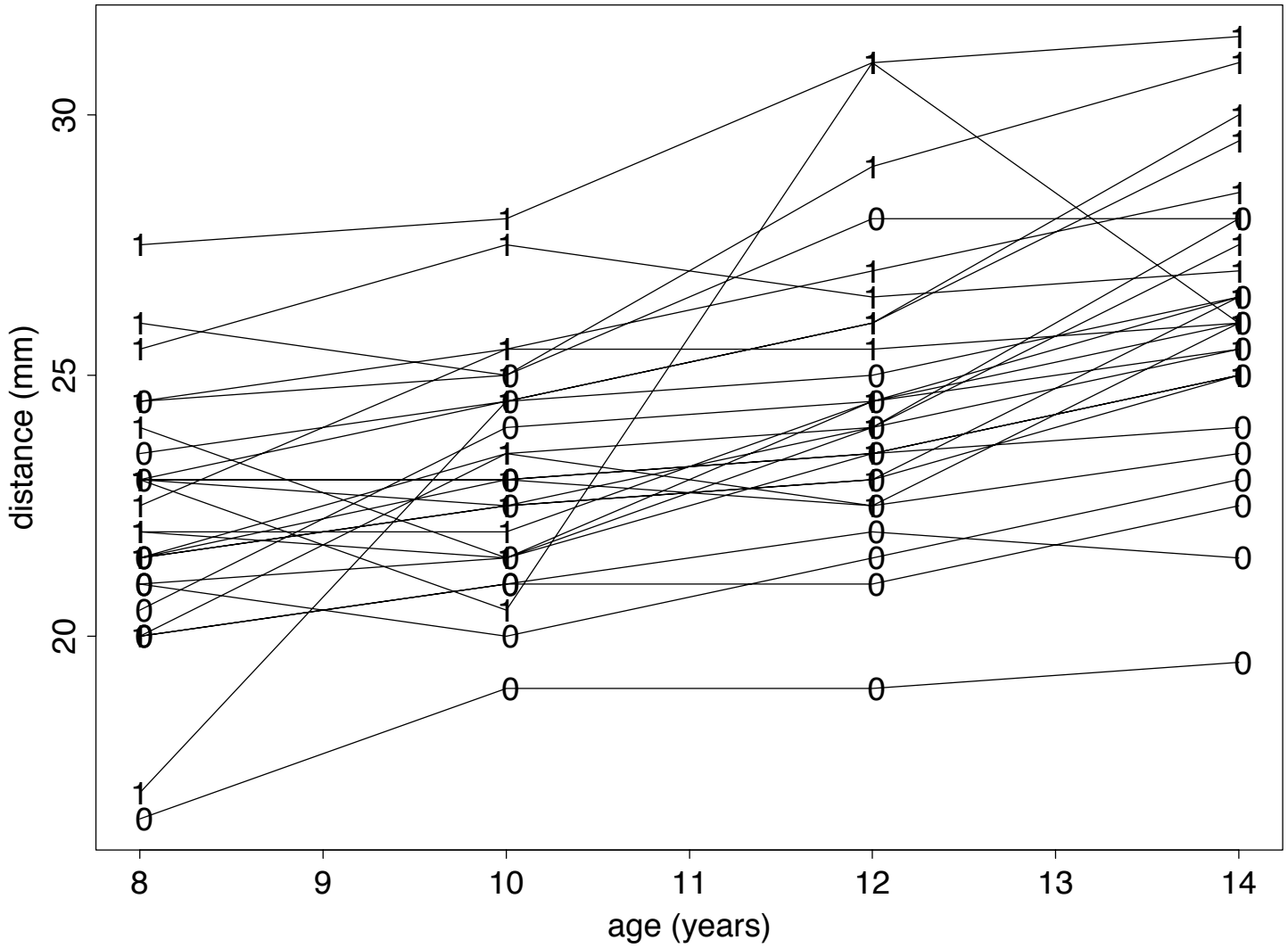
Questions of interest:

- Do things change over time?
- Understand pattern of change
- Is the pattern different for boys and girls? How?

Here is a plot of the data.

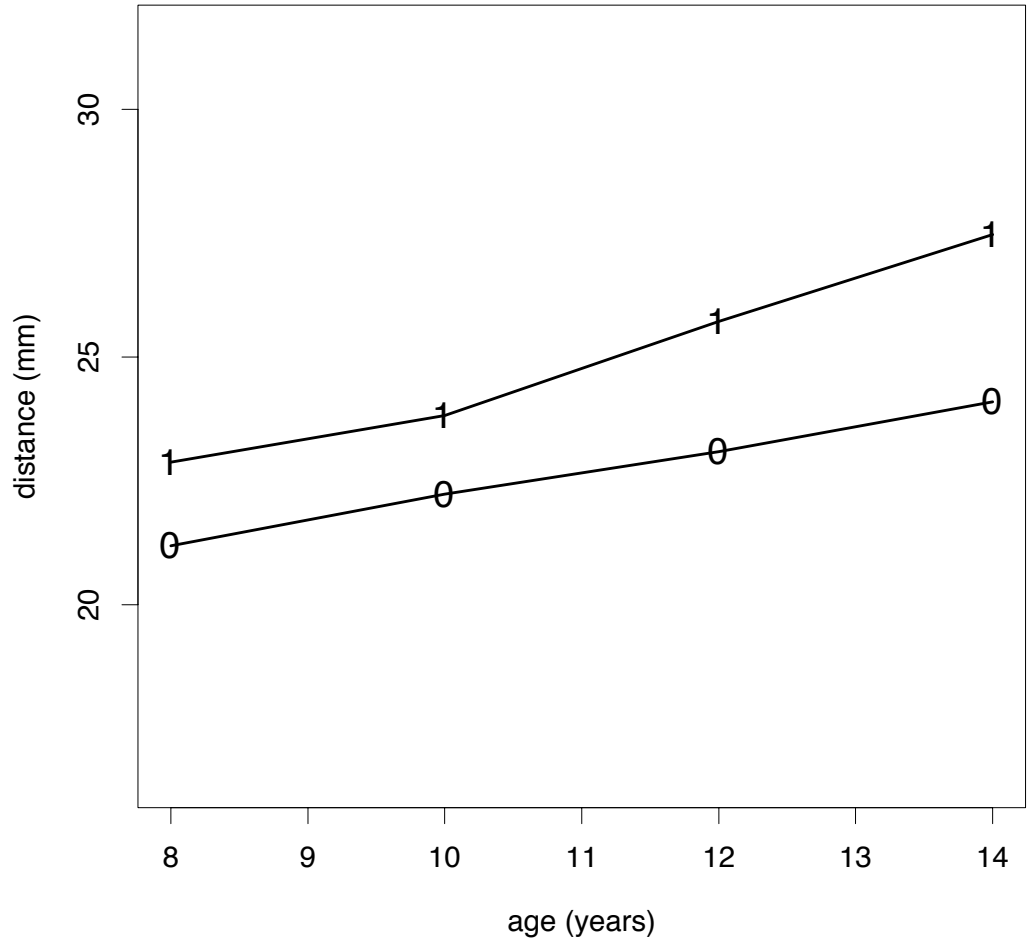
Dental Study Data

All data: 0=girl, 1 = boy



Dental Study Data

Sample average: 0= all girls, 1 = all boys



Remarks

- All children have all 4 measurements(no missing data, "balanced")
- Overall pattern of increasing distance measurements for boys and girls
- More specifically, the pattern for most children follows a rough straight line increase (with some "jitter")
- Average distance follows an approximate straight line pattern (although that for boys looks like it might curve. . .)
- The rate of change of the measurements with increasing age seems similar

To address these issues some formal model is required.

Multicenter AIDS Cohort Study (MACS) of HIV

The HIV virus destroys CD4+ cells (T-Lymphocytes, a vital component of the immune system) so that the number of CD4 cells in the blood of a patient will reduce after the subject is infected with HIV.

- A cohort of $m = 369$ men followed before and after HIV sero-conversion
- Interest on the natural history of HIV disease
- Important indicator of immune function is the CD4+ cell count:
  - collected on each subject approximately every six months
  - $n = 2376$ observations

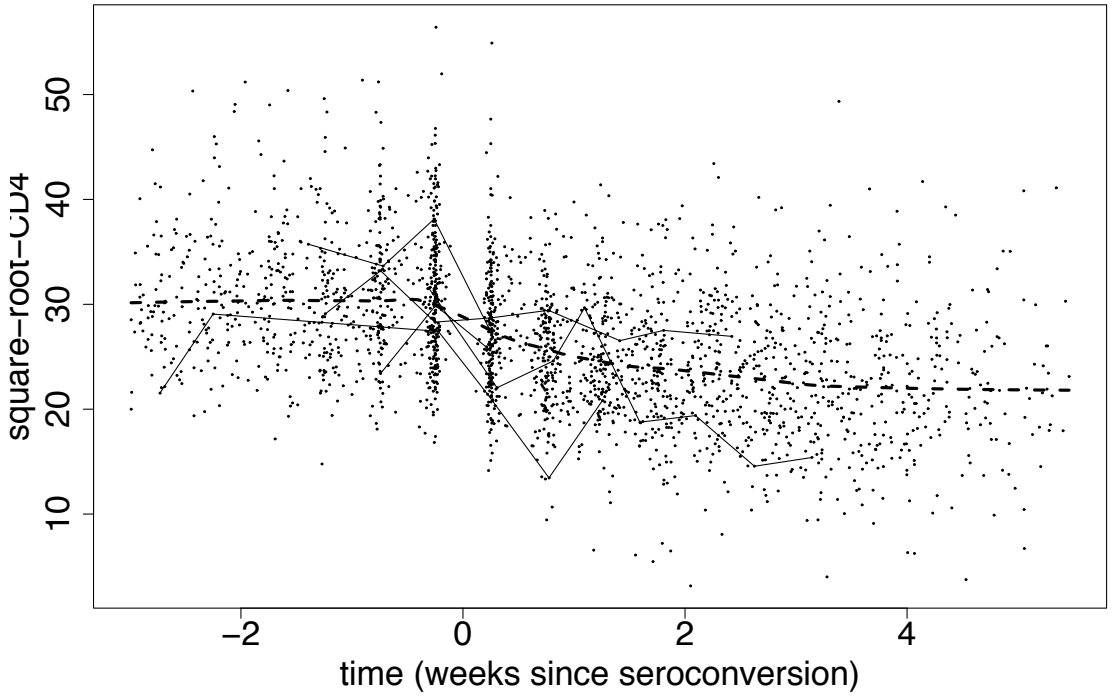
The measured variables were the *time* in weeks since seroconversion, *timedays* days since seroconversion, *cd4* CD4 Count, *age* (yrs) relative to arbitrary origin, *packs* of cigarettes smoked per day, *drugs* recreational drug use yes=1/no=0, *sexpart* number of sexual partners.

Scientific goals of the study

- characterize the typical time course of CD4+ cell depletion after HIV infection (natural history)
- characterize heterogeneity within and across men in CD4+ count and in progression of CD4+ depletion
- estimate time course (trajectory) of CD4+ count for individual men, accounting for substantial measurement error in CD4+ count
- study factors predicting levels and changes in CD4+ cell count

Here is a quick look at the data

MACS of HIV



Each observation is given by one point; Solid lines connect observations for four randomly-selected subjects (note high within-subject variability); Smooth dashed curve is the average time trend of CD4+ count.

Remarks

- lndividual CD4 profiles "jumparound" ("noisy") but many show a decrease over time
- Different subjects have CD4 measurements at different times (not "balanced")
- Some subjects drop out of the study, are administratively censored, or, worse, die hence no CD4 available
- Can't take averages (different time points)

Protein Content of Milk from Cows

These data comprise 19 weekly measurements of the protein content of milk samples from each of 79 Australian cows. Cows 1–25 were fed a barley diet, cows 26–52 a mixed diet of barley and lupins, and cows 53–79 a diet of lupins alone.

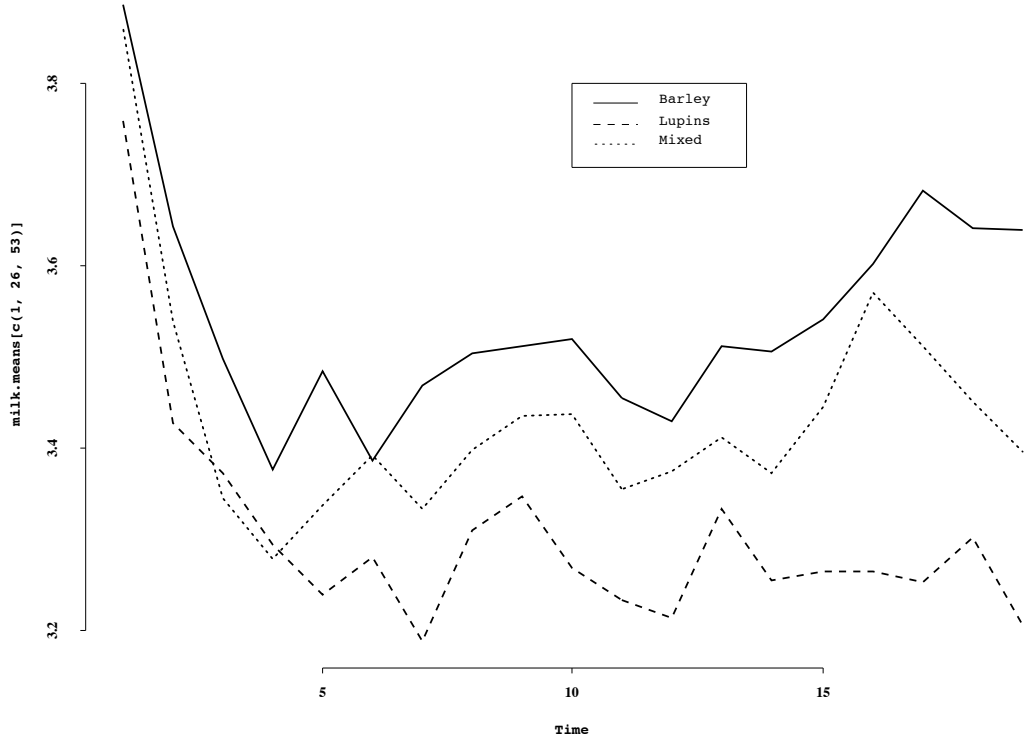
Time is measured in weeks since calving.

This dataset is notable because the experiment terminated at a fixed date, and so some of the time series are shorter than the others.

The goal is to study how diet affects the protein content of milk (treatment study): "response" is sequence of protein measurements.

Milk data

Mean protein content by time for each diet group (time=week since calving):



. . . suggests that barley substantially increases the protein content of milk.

Growth of Nepalese Children

- As part of a larger study, a cohort of $m = 200$ children was studied
- Anthropologic measurements were obtained from each child at $n_i = 5$ time points roughly 4 months apart
- Recording in these data were: the Childs ID, Childs age (months), Sex, Weight (kg), Height (cm), Arm circumference (cm), Current breastfeeding level, Nepali day of visit , Nepali month of visit, Nepali year of visit, Mothers age (yrs), Mothers literacy (1=yes), Number of Mother's children who died.

The scientific aims of the study are to:

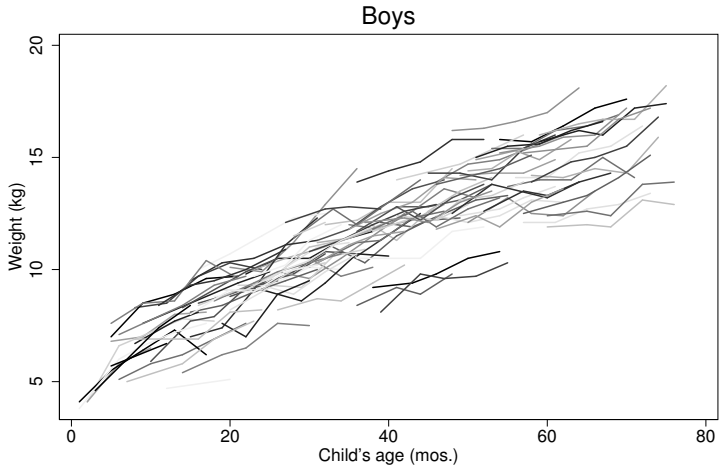
– characterize the growth patterns of Nepalese children (growth curve study)

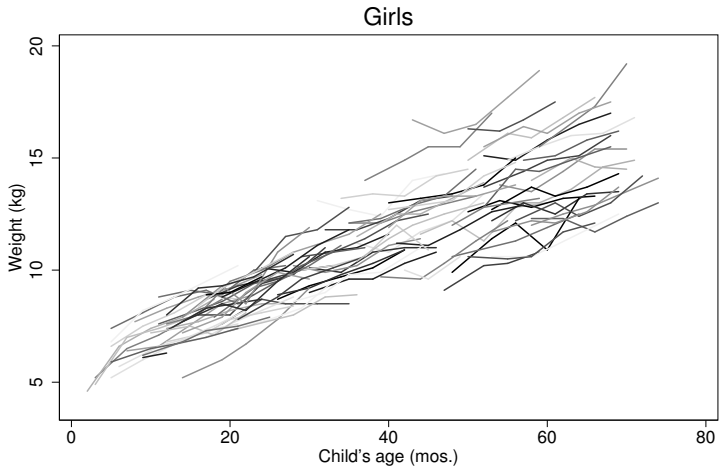
– describe the relationship between growth and child covariates and/or growth and maternal covariates (explanatory analysis)

Time scale is age (months) of the child

Nepalese Children (Weight)

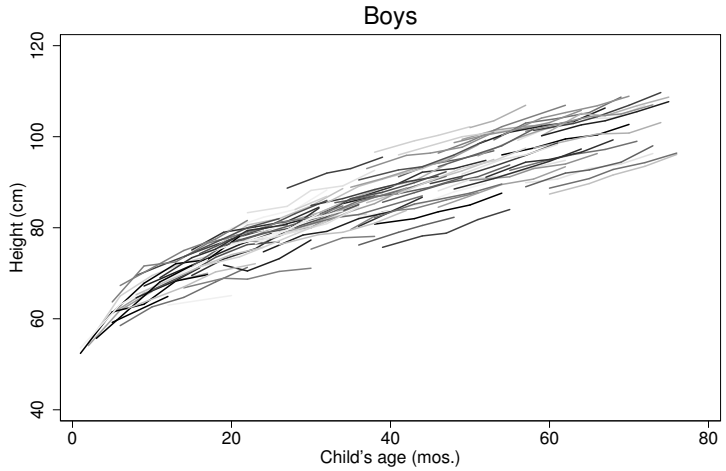
Here are the weight data for the boys and the girls:

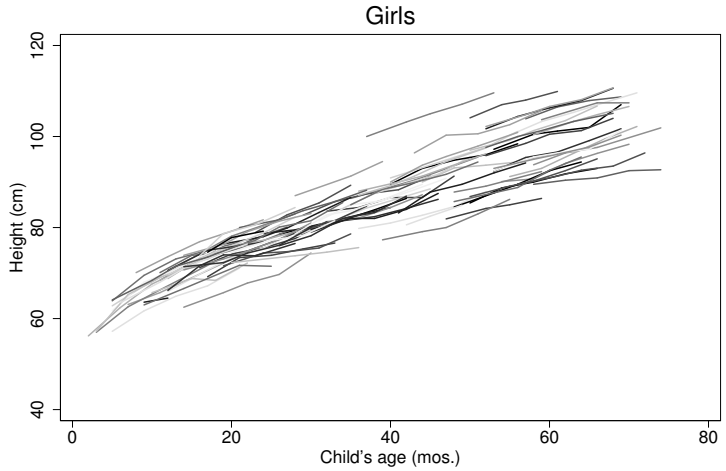




Nepalese Children (Height)

And here are the height data:





What Do These Examples Have in Common?

- Scientific objectives can be formulated as regression problems whose purpose is to describe the dependence of a response variable on explanatory variables;
- Repeated observations on each experimental unit:
  - Observations from one unit to the next are independent
  - Multiple observations on the same unit are dependent (correlated, associated)
- This correlation (association) makes longitudinal data powerful
- It also makes it a challenge to analyze (well)

Some Observations

- Complicated due to within-subject correlation due to variation over time within individuals or measurement error
- Emerging area (most research post-1980; most widespread software post-1990).
- Software still fairly new and not extremely stable.
- Lot of dust still to settle on best practice.

Goals for This Course

Describe the major ideas underlying longitudinal data analysis.

- Some of the basic parametric models and their properties.
- Estimation of parameters in basic models.
- Some exposure to inference and computing.

What Won't Be Achieved

- Typical graduate courses on longitudinal data analysis have 45-60 lectures, several assignments, computer labs, projects etc . . . These lectures (alone) will not train you comprehensively in the various nuances required for good applied longitudinal data analysis.
- Rather, you will be exposed (to varying degrees) to some of the basic ideas and principles of longitudinal data analysis.

Notation to be followed

- Y_{ij} = response variable and observed at time t_{ij} for observation $j = 1, \ldots, n_i$ on subject $i = 1, \ldots, m$.
- $\mathbf{x}_{ij}$ a vector of length p of explanatory variables observed at time t_{ij}
- $\mathbb{E}(Y_{ij}) = \mu_{ij}$, $\text{Var}(Y_{ij}) = v_{ij}$
- The set of repeated outcomes for unit i are collected into an n_i-vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$, with mean $\mathbb{E}(\mathbf{Y}_i) = \boldsymbol{\mu}_i$ and $n_i \times n_i$ covariance matrix $\text{Var}(\mathbf{Y}_i) = V_i$, $[V_i]_{jk} = \text{cov}(Y_{ij}, Y_{jk})$.

Most longitudinal analyses are based on a regression model

$$\mathbf{Y}_i = X_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i$$

where X_i is a $n_i \times p$ matrix with $\mathbf{x}_{ij}$ in the jth row and $\boldsymbol{\epsilon}_i$ a vector of random errors.

Books

- DIGGLE, P.J., HEAGERTY, P., LIANG, K.Y. AND ZEGER, S.L. (2002). *Analysis of Longitudinal Data*(second edition). Oxford: Oxford University Press.
- FITZMAURICE, G.M., LAIRD, N.M. AND WARE, J.H. (2004). *Applied Longitudinal Analysis*. New Jersey: Wiley.
- PINEIRO, J.C. AND BATES, D.M. (2000) *Mixed-Effects Models in S and S-Plus*, Springer-Verlag, New York.
- VERBEKE, G. AND MOLENBERGHS, G. (2000) *Linear mixed models for longitudinal data*. Springer series in statistics, Springer, New York.
- MCCULLOGH C.E AND SEARLE, S.R. (2001) *Generalized, linear and mixed models*. Wiley, New York.

Exploring longitudinal data

Questions to be answered:

- What is nature of response variable? Continuous? Count? Binary?
- is the data balanced or unbalanced?
- What is the degree of (unintended) incomplete or missing data relative to the design?
- What is the distribution of responses by time (time plots):
  - box-plot or similar for highly balanced settings
  - scatter plot for unbalanced settings
- What is the population average or mean trend with time?
- add curves to time plots
- explore correlation structure

Balanced and unbalanced data

- balanced design : intended measurement times common to all subjects
- balanced data : actual measurement times common to all subjects

When data are unbalanced, an important question is **WHY**:

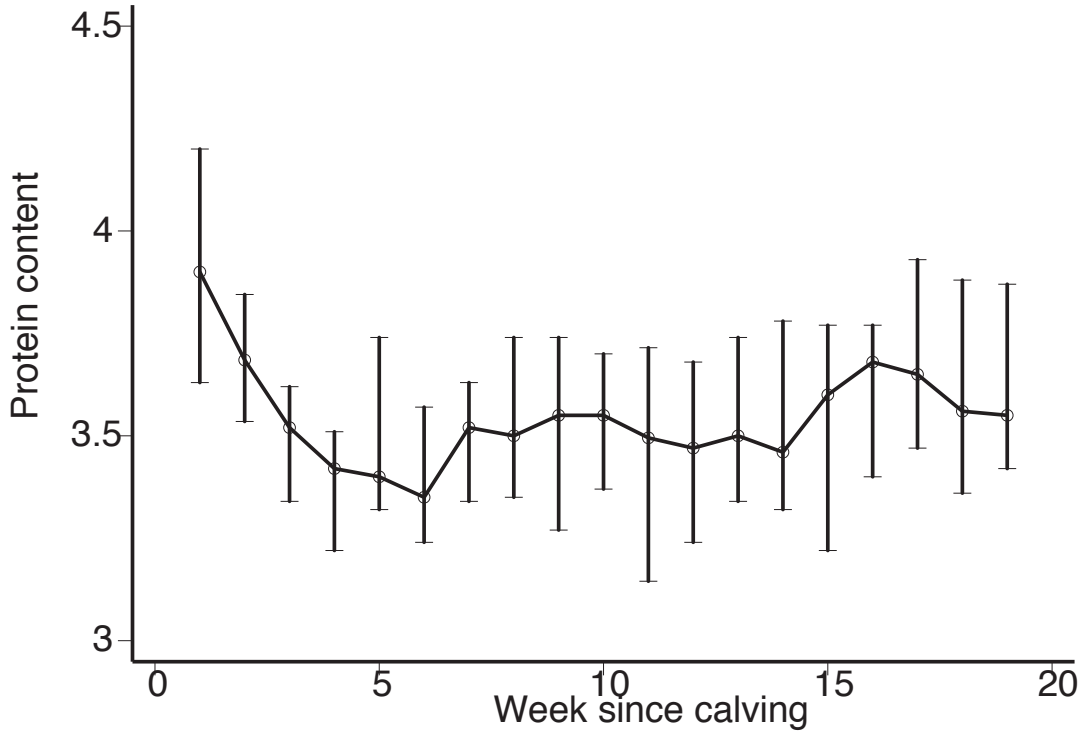
- by design?
- missing values?
- observational data?

Graphical summaries

Standard graphical summaries:

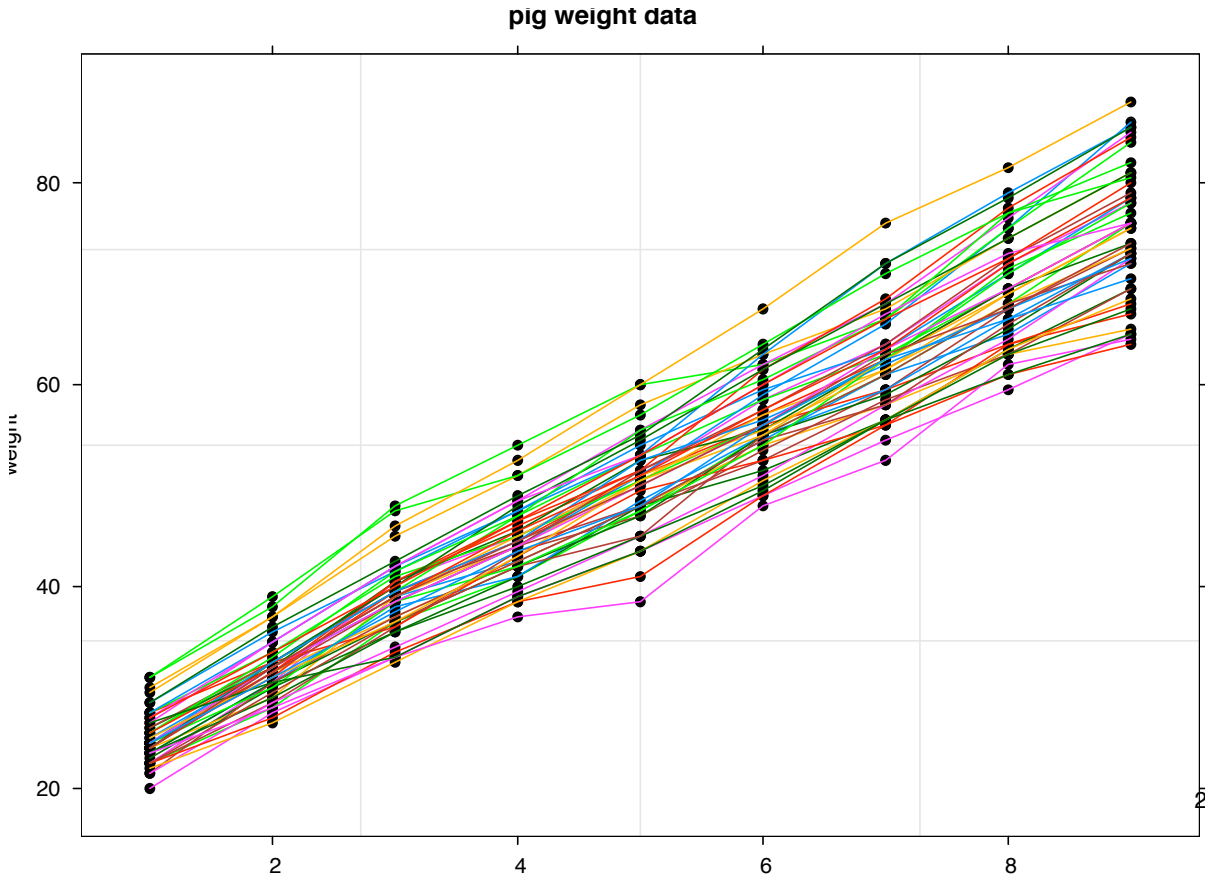
- plot of measurements against time
  - error-plots (to present (features of) the distribution of response Y across time)
  - line-plots: connected line segments for each subject (spaghetti plots)
  - Add a nonparametric smooth curve as preliminary estimate of mean response (fitting smooth curves to longitudinal data)
- Decompose the data into cross-sectional patterns and longitudinal patterns
- for balanced design display a scatterplot matrix

Example: milk data - Median (IQ range) for barley diet

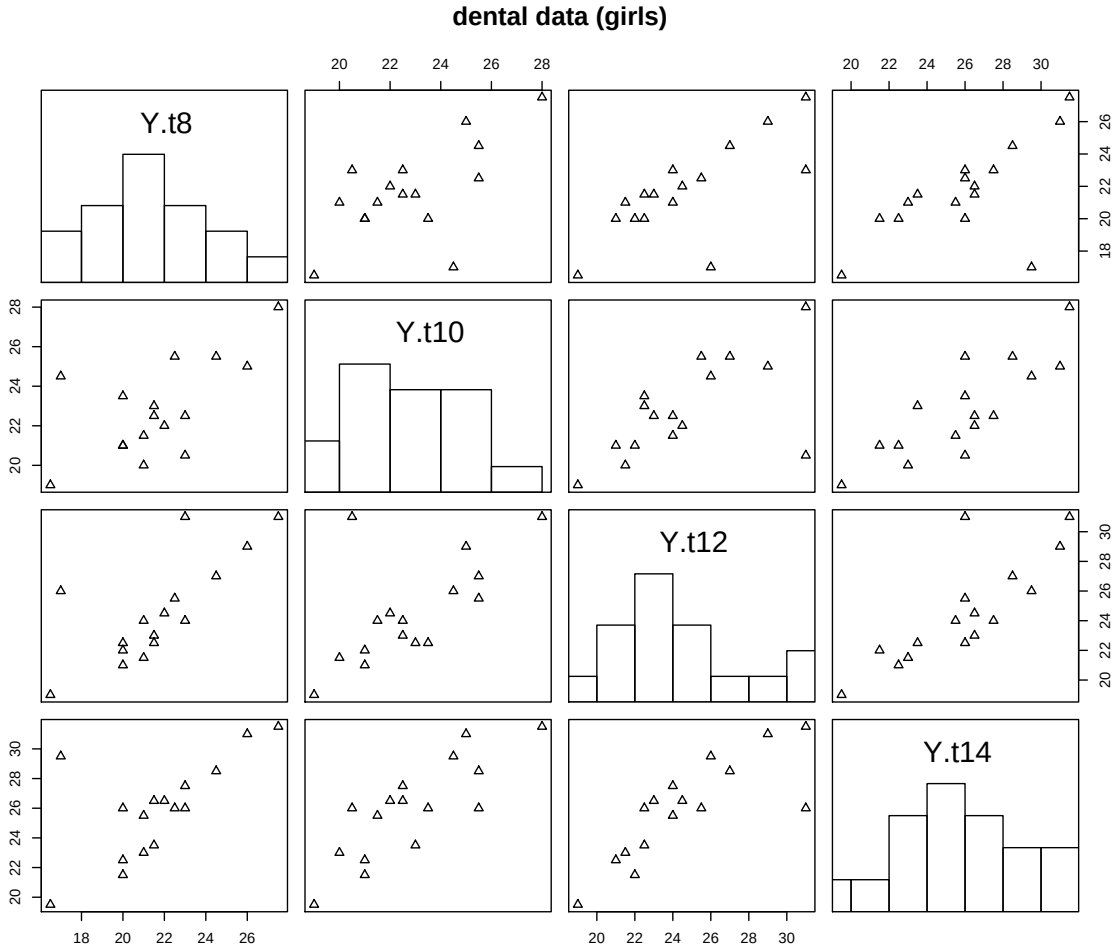


Such plots are easier to read than box-plots when there are many time points.

Pig weights – line-plot



Dental data: scatterplot matrix



Exploring mean response profiles by smooth curve fitting

Nonparametric regression models can be used to estimate the mean response profile as a function of time.

- Data (y_i, t_i) $i = 1, \ldots, n$

We want to fit an unknown mean response curve $\mu(t)$ in the underlying model

$$Y_i = \mu(t_i) + \epsilon_i$$

- Kernel estimation
- Smoothing Spline
- Local polynomial smoothing

Kernel Estimation (local running mean)

- Selection of window of width h centered at time t;
- $\hat{\mu}(t)$ is the average of Y values of all points which are visible in that window
- To obtain an estimator of the smooth curve at every time, slide a window from the extreme left to the extreme right, calculating the average of the points within the window every time
- Weighted local running mean uses a weighting function that changes smoothly with time and gives stronger weights to the observations closer to t (e.g. Gaussian kernel $K(u) = \exp(-0.5u^2)$).

Smoothing spline

- Is the "smooth" function $s(t)$ which minimizes the criterion

$$J(\lambda) = \sum_{i=1}^{n} \{y_i - s(t_i)\}^2 + \lambda \int s''(t)^2 dt$$

- $s(t)$ minimizes the criterion if and only if it is a piecewise cubic polynomial (a natural cubic spline with knots at each time t_i.
- Such methods may be easily seen as projection regression methods on specific functional bases.

Local polynomial smoothing

- At each time point x_0, one fits a polynomial

$$P_{x_0}(x) = \beta_0 + \beta_1(x - x_0) + \cdots + \beta_p(x - x_0)^p$$

  of degree p, by weighted least squares with weights $w_i = K_h(x_i - x_0)$.

- The estimation $\hat{\mu}(x_0)$ is given by $P_{x_0}(x_0) = \beta_0$.

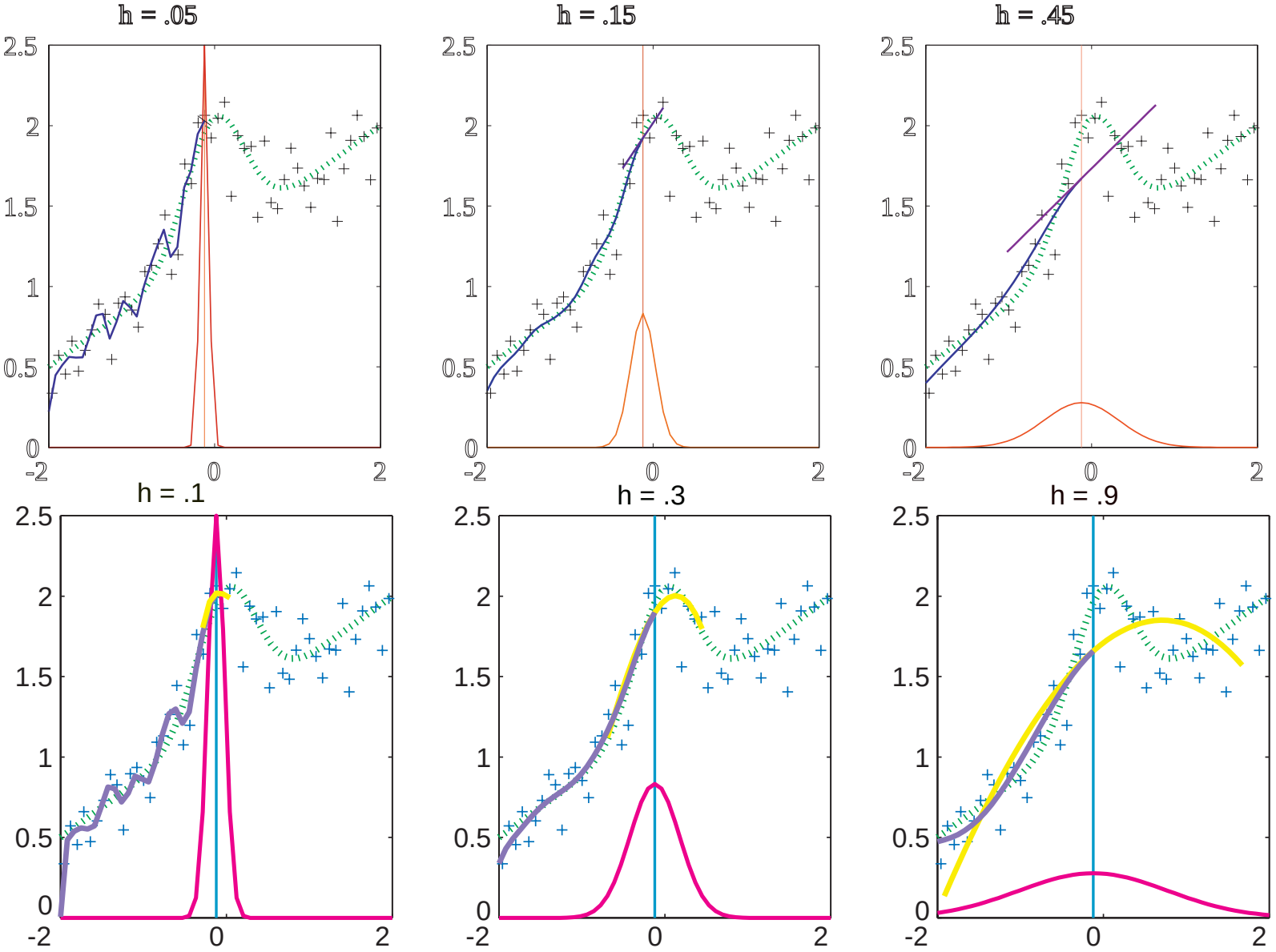
In matrix notation:

$$\hat{m}(x) = \mathbf{e}_1^T \left(\mathbf{X}_{x_0}^T \mathbf{W}_{x_0} \mathbf{X}_{x_0} \right)^{-1} \mathbf{X}_{x_0}^T \mathbf{W}_{x_0} \mathbf{Y},$$

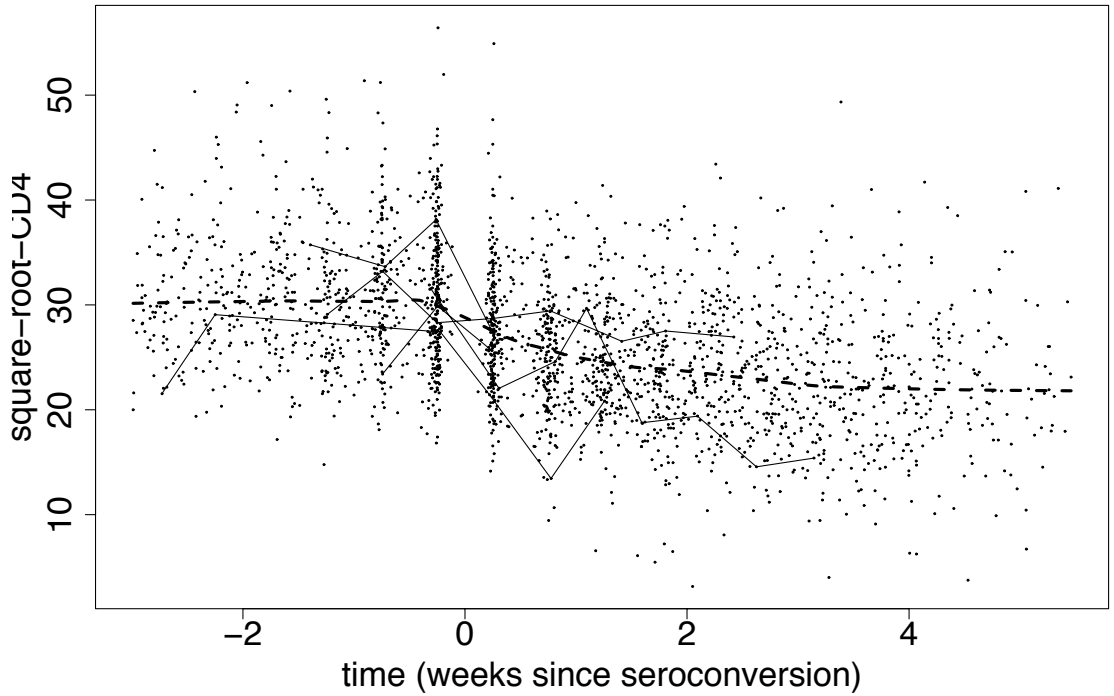
where $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$, $\mathbf{Y} = (y_1, \ldots, y_n)^T$, $\mathbf{W}_{x_0} = \text{diag}\{K_h(x_i - x_0)\}$ et

$$\mathbf{X}_{x_0} = \begin{pmatrix} 1 & x_1 - x_0 & \cdots & (x_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & x_n - x_0 & \cdots & (x_n - x_0)^p \end{pmatrix}.$$

Example



Example : MACS of HIV



The smooth dashed curve is the mean time trend of CD4+ count obtained using a local polynomial method.

Remarks

There are many omitted details in the above discussion:

- choice of weighting function (kernel)
- what to do at the edges of the data
- reducing effects of outliers
- choice of bandwidth or smoothing parameter

Graphical methods to separate CS and Longitudinal Patterns

- $Y_{ij} = \beta_0 + \beta_C \bar{x}_i + \beta_L (x_{ij} - \bar{x}_i) + \epsilon_{ij},\ i = 1, \ldots, m; j = 1, \ldots, n.$

This model implies two facts:

1. $\bar{Y}_i = \beta_0 + \beta_C \bar{x}_i + \bar{\epsilon}_i,\ i = 1, \ldots, m$, capturing cross-sectional effects parameterized by β_C
2. $Y_{ij} - \bar{Y}_i = \beta_L (x_{ij} - \bar{x}_i) + \epsilon_{ij} - \bar{\epsilon}_i$ capturing longitudinal effects parameterized by β_L

This suggests that β_C can be investigated by plotting $\bar{Y}_i$ versus $\bar{x}_i$, and β_L can be investigated by plotting $Y_{ij} - \bar{Y}_i$ versus $x_{ij} - \bar{x}_i$.

Example: Arm circumference and weight in Nepalese girls

Suppose we wish to investigate the degree to which arm circumference reflects weight differences:

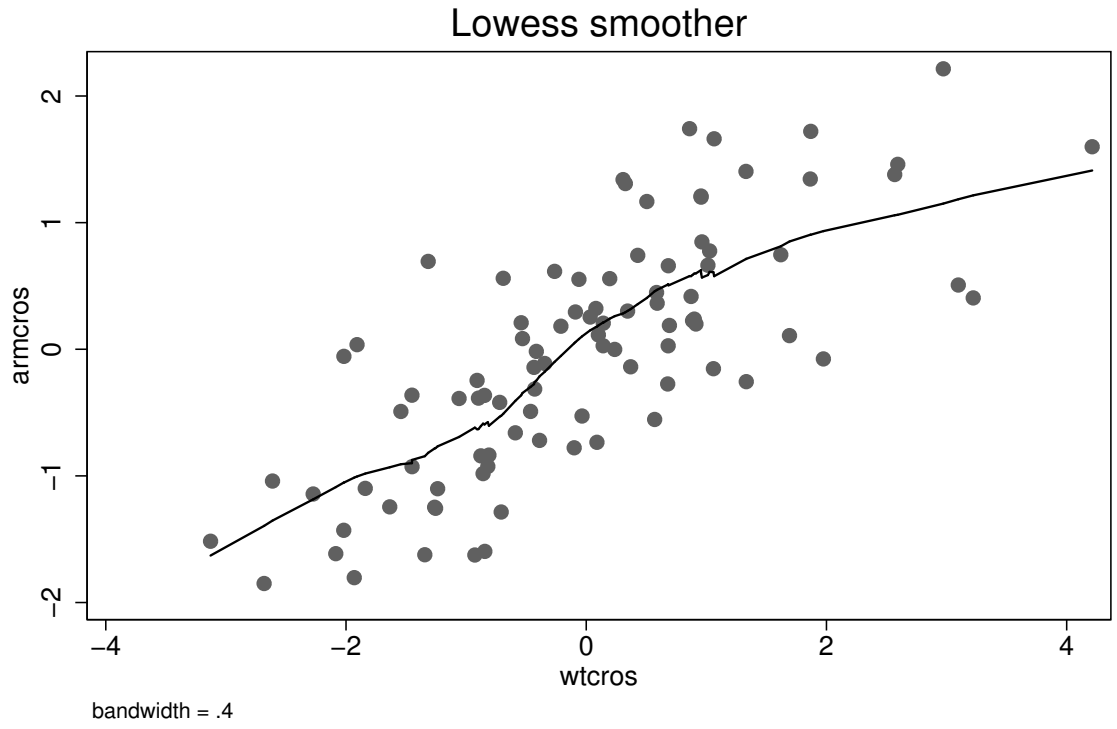
- Are differences in average weight across girls reflected in differences in average arm circumference across girls? (cross-sectional question)
- Are changes in weight for a given girl reflected in changes in arm circumference? (longitudinal question)

A (working) model for this question is $Y_{ij} = \beta_0 + \beta_C \bar{x}_i + \beta_L (x_{ij} - \bar{x}_i) + \epsilon_{ij}$
where: Y_{ij}= arm circumference with age-effect removed
x_{ij} = weight with age-effect removed
We can use a smoothing model fit to remove the age trends in arm circumference and in weight by first fitting the models

$$arm = \mu_1(age) + \epsilon_1 \quad wt = \mu_2(age) + \epsilon_2$$

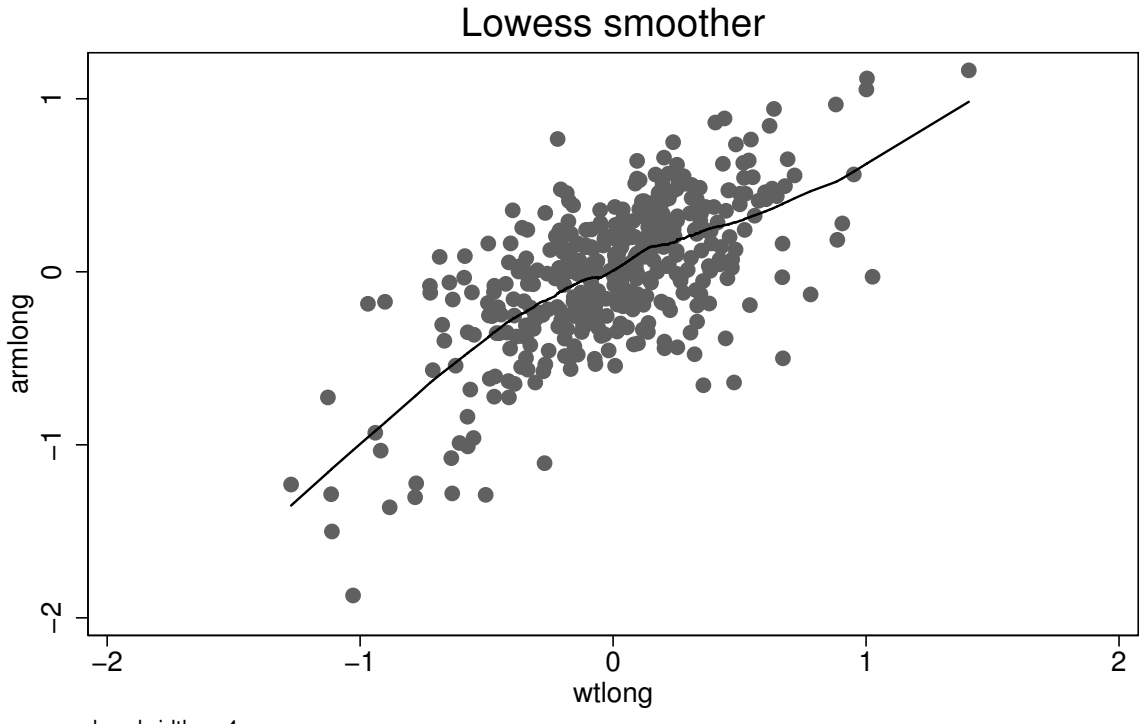
Example (Cont.)

To investigate these questions, we examine the residuals after removing the effects of age, `wtresid` and `armresid`



Mean arm circumference $\bar{Y}_i$ versus mean weight $\bar{x}_i$ (age effect removed).

Example (Cont.)



Change in arm circumference $Y_{ij} - \bar{Y}_i$ versus change in weight $x_{ij} - \bar{x}_i$ (age effect removed).

Conclusion: Both between-girl differences **and** within-girl differences in weight are strongly reflected in arm circumference.

Exploring correlation structure: the variogram

The *variogram* of a stationary random process $Y(t)$:

$$V(u) = \frac{1}{2}\text{Var}\{Y(t) - Y(t-u)\}$$

If $\mathbb{E}[Y(t)] = 0$ and $\text{Cov}\{Y(t), Y(t-u)\} = \gamma(u)$, it is easy to see that

$$V(u) = \gamma(0) - \gamma(u).$$

So why bother with the variogram?

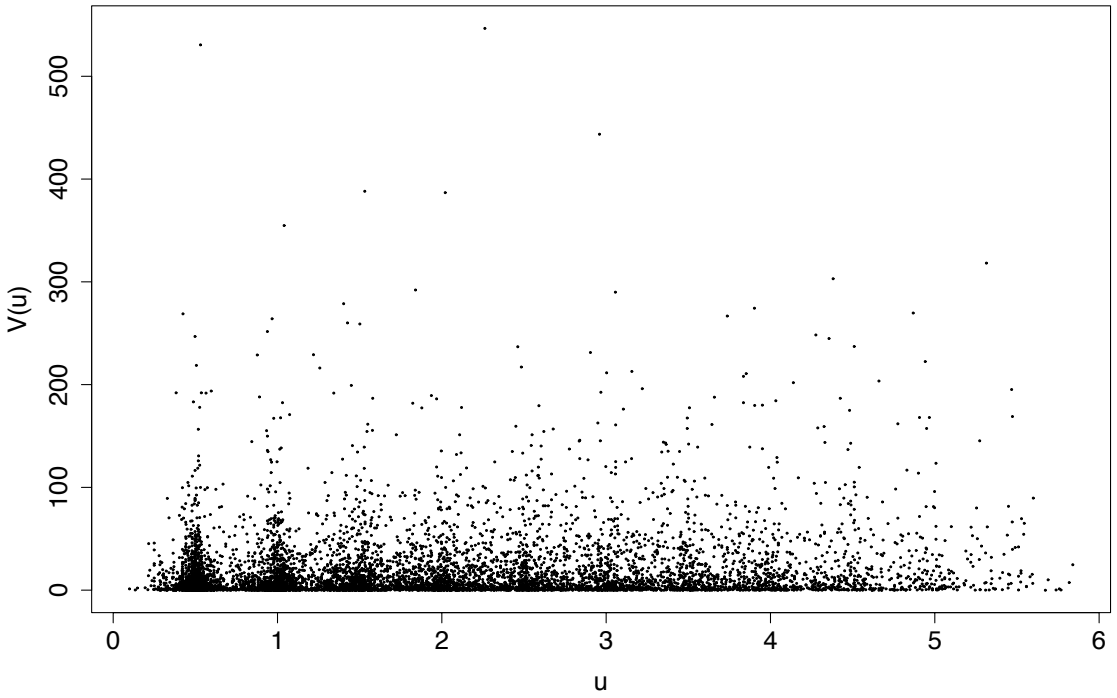
- it is also well-defined for some non-stationary processes (useful if we want a diagnostic for non-stationarity)
- it is easier to estimate from irregularly spaced data.

Estimating the variogram

Let r_{ij} = residual from preliminary model for mean response

- Define
$$v_{ijk\ell} = \frac{1}{2}(r_{ij} - r_{k\ell})^2$$
- Calculate
$\hat{V}(u)$ = average of all quantities $v_{iji\ell}$ such that $|t_{ij} - t_{i\ell}| \simeq u$
- Estimate process variance by
$\hat{\sigma}^2$ = average of all quantities $v_{ijk\ell}$ such that $i \neq k$.

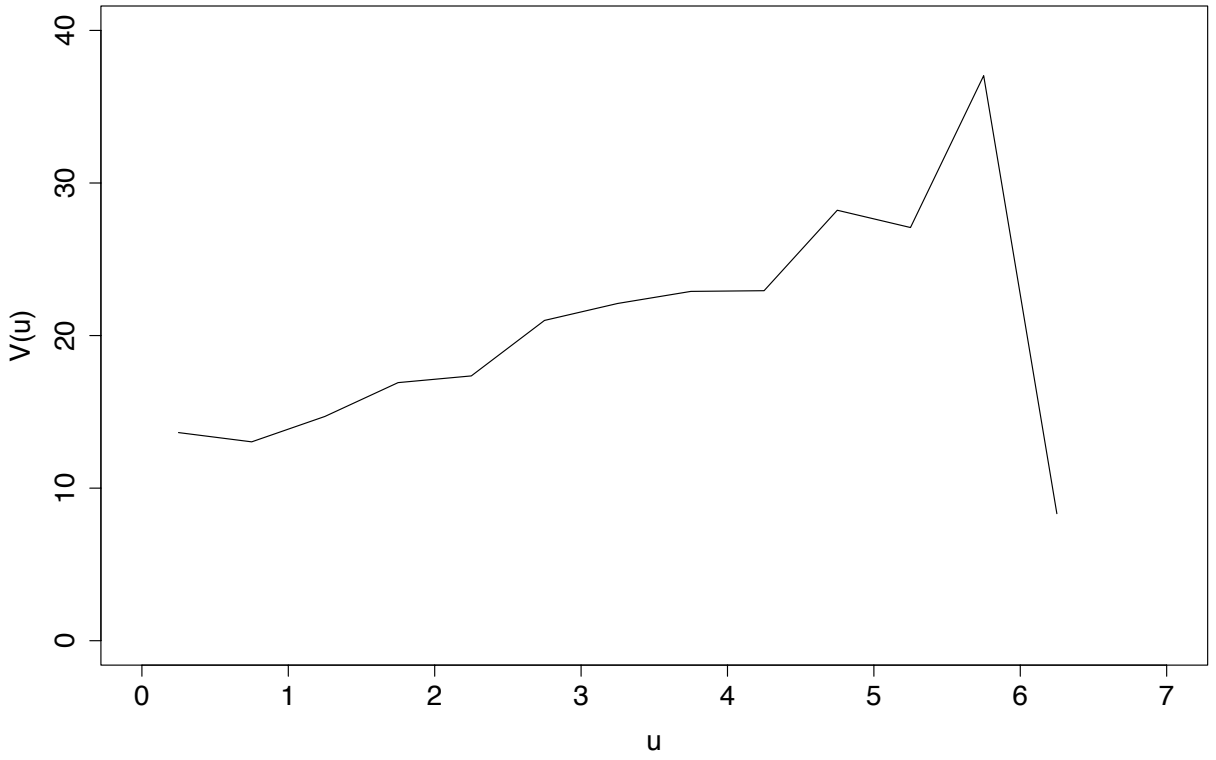
Example : Square-root CD4+ cell numbers



Note the very large sampling fluctuations.

Smoothing the empirical variogram

- For irregularly spaced data, group time-differences u into bands and take averages of corresponding $v_{ijk\ell}$.
- For data from a balanced design, usually no need to average over bands of values for u.



Linear models for longitudinal data

$$\mathbb{E}(Y_{ij}) = x_{ij1}\beta_1 + \cdots + x_{ijp}\beta_p, \quad i = 1, \ldots, m;\ j = 1, \ldots, n_i$$

$$\mathbf{Y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, m.$$

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Within-subject residual sequences $\boldsymbol{\epsilon}_i$ are typically correlated. Modelling the correlation is important to be able to obtain correct inferences on regression coefficients $\boldsymbol{\beta}$.

Three basic elements of correlation structure:

- random effects
- autocorrelation or serial dependence
- noise, measurement error

Estimation is achieved via *weighted least squares*.

Some simple models

Consider the pig weight data introduced before. Let Y_{ij} denote the weight of pig i on week j and let $x_j = j$ be the corresponding week number. The simplest model is to treat the data cross-sectionally using an ordinary least squares model

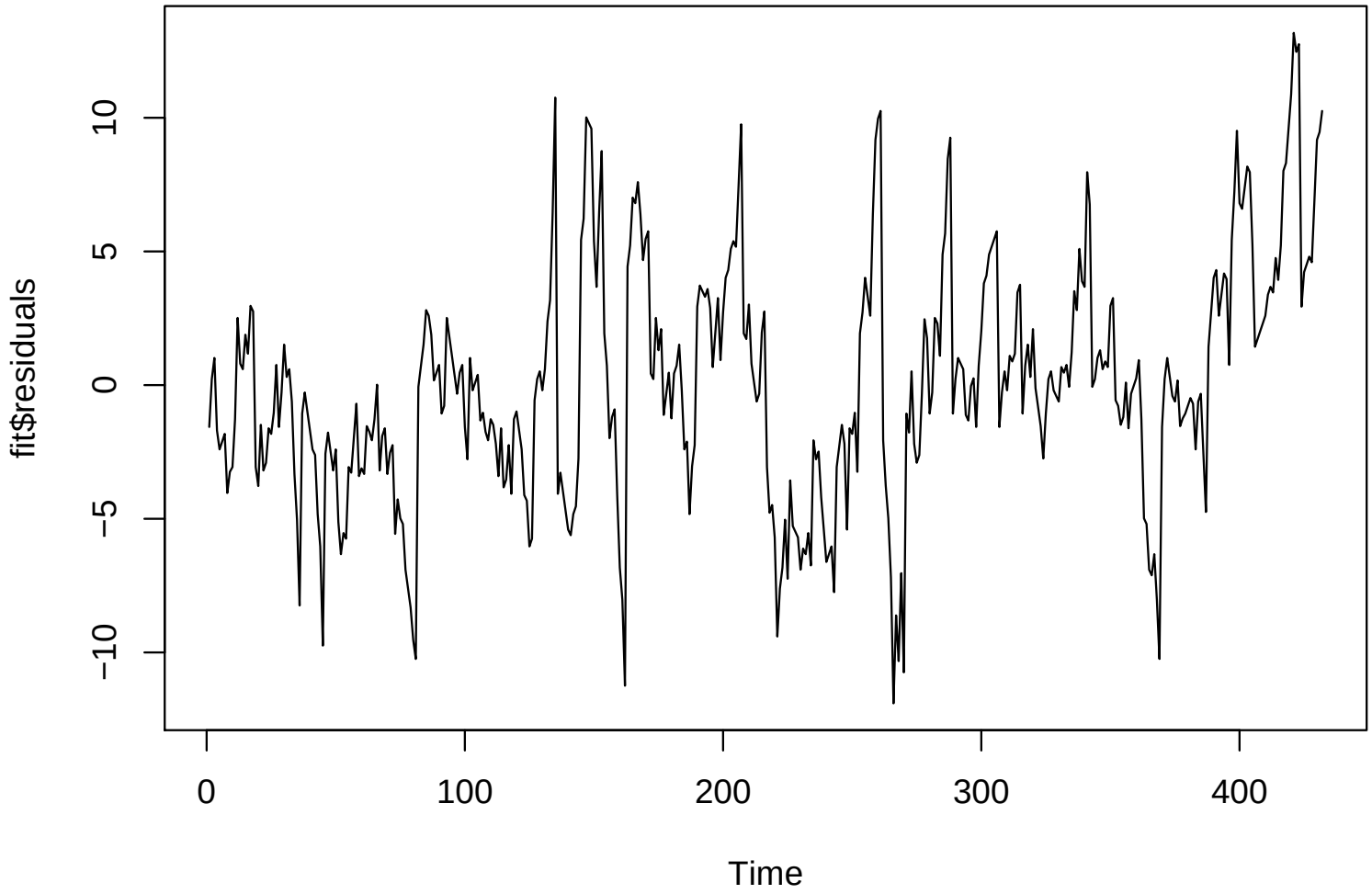
$$Y_{ij} = \beta_0 + \beta_1 x_j + \epsilon_{ij}, \quad 1 \leq i \leq 48, j = 1, \ldots, 9$$

with ϵ_{ij} i.i.d. $N(0, \sigma_\epsilon^2)$ which leads to a slope estimate

$$\hat{\beta}_1 = 6.21, \quad \widehat{\text{st.dev}}(\hat{\beta}_1) = 0.0818$$

But there are problems. Inspection of the line-plot shows that the scatterplot for each individual pig is less variable, so using a within-pig information should be beneficial. Moreover the previous model ignores the correlation of measurements pertaining to the same pig.

Pattern of residuals from naive fit



Pig weights (cont.)

Since the slopes look about the same but each pig seems to have his/her own intercept a remedy could be to fit

$$Y_{ij} = \beta_{0i} + \beta_1 x_j + \epsilon_{ij}$$

for $1 \leq i \leq 48$ with $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$.

But this model has 50 parameters and only β_1 and σ_ϵ^2 are interpretable. Moreover, it gives too much credence to the pigs used in the study.

Random Intercept Model

A better model is:

$$Y_{ij} = U_i + \beta_0 + \beta_1 x_{ij} + \epsilon_{ij},$$

where

$$U_i \text{ are i.i.d. } \sim N(0, \sigma_U^2)$$

and independent of the

$$\epsilon_{ij} \text{ i.i.d. } \sim N(0, \sigma_\epsilon^2)$$

which falls into the class of *compound symmetry* models

$$Y_{ij} - \mu_{ij} = U_i + \epsilon_{ij}, \quad U_i \sim N(0, \sigma_U^2), \epsilon_{ij} \sim N(0, \sigma_\epsilon^2).$$

Within subject covariance

Let us compute the covariance between different measures, say $j \neq j'$, of the same subject i:

$$\begin{aligned}
\text{Cov}(Y_{ij}, Y_{ij'}) &= \text{Cov}(\mu_{ij} + U_i + \epsilon_{ij}, \mu_{ij'} + U_i + \epsilon_{ij'}) = \text{Cov}(U_i + \epsilon_{ij}, U_i + \epsilon_{ij'}) \\
&= \text{Cov}(U_i, U_i) + \text{Cov}(U_i, \epsilon_{ij'}) + \text{Cov}(\epsilon_{ij}, U_i) + \text{Cov}(\epsilon_{ij}, \epsilon_{ij'}) \\
&= \sigma_U^2 + 0 + 0 + 0 = \sigma_U^2
\end{aligned}$$

For $j = j'$, $\text{Cov}(Y_{ij}, Y_{ij'}) = \text{Var}(U_i + \epsilon_{ij}) = \sigma_U^2 + \sigma_\epsilon^2$

For $i \neq i'$, $\text{Cov}(Y_{ij}, Y_{i'j'}) = 0$.

Concrete example

$m = 3, n_1 = 2, n_2 = 3, n_3 = 2$. Under the compound symmetry model the covariance matrix of the vector

$$\mathbf{Y} = (Y_{11}, Y_{12}, Y_{21}, Y_{22}, Y_{23}, Y_{31}, Y_{32})'$$

is

$$\begin{bmatrix} \begin{pmatrix} \sigma_U^2 + \sigma_\epsilon^2 & \sigma_U^2 \\ \sigma_U^2 & \sigma_U^2 + \sigma_\epsilon^2 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} \sigma_U^2 + \sigma_\epsilon^2 & \sigma_U^2 & \sigma_U^2 \\ \sigma_U^2 & \sigma_U^2 + \sigma_\epsilon^2 & \sigma_U^2 \\ \sigma_U^2 & \sigma_U^2 & \sigma_U^2 + \sigma_\epsilon^2 \end{pmatrix} & \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} \sigma_U^2 + \sigma_\epsilon^2 & \sigma_U^2 \\ \sigma_U^2 & \sigma_U^2 + \sigma_\epsilon^2 \end{pmatrix} \end{bmatrix}$$

Remarks on Random Intercept Model

- Invokes a positive (between measurements of same subject) within-subject correlation

$$\rho = \text{Corr}(Y_{ij}, Y_{ik}) = \sigma_U^2 / (\sigma_U^2 + \sigma_\epsilon^2)$$

- Correlation is same for all subjects and regardless of distance apart in time (disadvantage).
- This type of correlation structure is known as *exchangeable correlation* or *compound symmetry*.
- The β_0 and β_1 (and more generally μ_{ij}) are called *fixed effects*. The U_i are called *random effects*.
- Since the model contains both fixed and random effects it is a special case of a *mixed effects model* or mixed model for short.
- The parameters σ_U^2 and σ_ϵ^2 are often referred to as *variance components*.

Estimation of model parameters

- Review of linear regression model
- Ordinary least squares estimation
- General linear model with correlated errors
- Weighted least squares
- Maximum Likelihood Estimation (MLE) and Restricted maximum Likelihood Estimation (RMLE)

Linear regression model

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

- X is a $n \times p$ matrix,
- $\boldsymbol{\beta} \in \mathbb{R}^p$
- $\mathbb{E}(\boldsymbol{\epsilon}) = 0$ and $\text{Var}(\boldsymbol{\epsilon}) = \sigma_\epsilon^2 I_n$.
- Often $\boldsymbol{\epsilon} \sim MVN(0, \sigma_\epsilon^2 I_n)$.

When $x_{i1} = 1$ for all i, then β_1 is the intercept.

This is the model used for fitting naively the pig weight data.

To estimate the unknown parameters one uses ordinary least squares (OLS) in absence of distributional assumptions or maximum likelihood otherwise.

Ordinary least squares

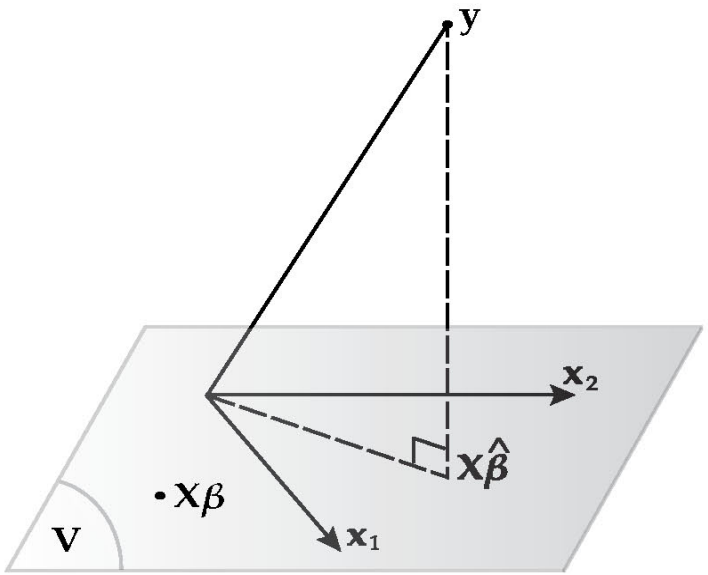
The ordinary least squares estimator of $\boldsymbol{\beta}$ is the estimator obtained by minimizing with respect to $\beta_1, \ldots, \beta_p$ the expression:

$$SS(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i - x_{i1}\beta_1 - x_{i2}\beta_2 - \ldots x_{ip}\beta_p \right]^2 = \|\mathbf{y} - X\boldsymbol{\beta}\|^2 .$$

Choose $\hat{\boldsymbol{\beta}}$ so that the distance from $\mathbf{Y}$ to $X\hat{\boldsymbol{\beta}}$ is as small as possible !

Geometry of ordinary least squares

Let V be the vector space spanned by the columns of the design matrix X. Geometrically one has



Geometrical interpretation of least squares for a linear model with two regressors. The space V is the plane of $\mathbb{R}^n$ spanned by the columns $\mathbf{x}_1$ and $\mathbf{x}_2$ of the design matrix X. The estimation is obtained as the orthogonal projection $X\hat{\boldsymbol{\beta}}$ of the observed vector $\mathbf{y}$ onto this plane.

The OLS estimators

$\hat{\boldsymbol{\beta}}$ that minimizes $SS(\boldsymbol{\beta})$ makes the residual vector $\mathbf{Y} - X\hat{\boldsymbol{\beta}}$ orthogonal to the space V spanned by the columns of X.

- $\hat{\boldsymbol{\beta}} = \left({}^tXX\right)^{-1}\,{}^tX\mathbf{Y} \equiv D\mathbf{Y}$
- $\hat{\mathbf{Y}} = X\left({}^tXX\right)^{-1}\,{}^tX\mathbf{Y} = H\mathbf{Y}$
- $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$
- $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma_\epsilon^2\left({}^tXX\right)^{-1}$
- $\hat{\boldsymbol{\beta}}$ is of minimum variance among all other unbiased estimators of $\boldsymbol{\beta}$ that are linear in $\mathbf{Y}$ (Gauss-Markov).

Distributional theory

When $\boldsymbol{\epsilon} \sim MVN(0, \sigma^2 I_n)$ then the log-likelihood is

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}{}^t(\mathbf{y} - X\boldsymbol{\beta})(\mathbf{y} - X\boldsymbol{\beta}) - \frac{n}{2}\ln(2\pi).$$

One can also write it :

$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}SS(\boldsymbol{\beta}) - \frac{n}{2}\ln(2\pi)$$

$$\sup_{(\boldsymbol{\beta},\sigma^2)} \mathcal{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = \sup_{\sigma^2} \sup_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}, \sigma^2; \mathbf{y}).$$

Hence the OLS estimator of $\boldsymbol{\beta}$ is also the MLE of $\boldsymbol{\beta}$ and the MLE of σ^2_ϵ is $\tilde{\sigma}^2 = \sigma^2(\hat{\boldsymbol{\beta}}) = SS(\hat{\boldsymbol{\beta}})/n$. It can be easily shown (see later) that the RMLE of σ^2_ϵ is $\hat{\sigma}^2 = SS(\hat{\boldsymbol{\beta}})/(n-2)$.

Moreover, $\hat{\boldsymbol{\beta}} \sim MVN(\boldsymbol{\beta}, \sigma^2_\epsilon DD')$ and $\hat{\boldsymbol{\beta}}$ is of minimum variance among all unbiased estimators of $\boldsymbol{\beta}$.

Remarks

With correlated errors, $\text{Var}(\mathbf{Y}) = V$. In the general linear model, and if one focus on the population mean (main interest on $\boldsymbol{\beta}$ and a simple structure for V), one could still use OLS (in a more clever way) to estimate $\boldsymbol{\beta}$ by proceeding as follows:

- $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}; \mathbb{E}(\boldsymbol{\epsilon}) = 0; \text{Var}(\boldsymbol{\epsilon}) = V.$
- $\hat{\boldsymbol{\beta}} = \left({}^tXX\right)^{-1} {}^tX\mathbf{Y} \equiv D\mathbf{Y}$
- $\text{Var}(\hat{\boldsymbol{\beta}}) = DVD' \simeq D\hat{V}D'$, with $\hat{V}$ sample covariance matrix of ordinary residuals.
- Under the Gaussian assumption $\hat{\boldsymbol{\beta}} \sim MVN(\boldsymbol{\beta}, D\hat{V}D')$.

Remarks (cont.)

Good points:

- technically simple, often reasonably efficient
- don't need to specify covariance structure

Bad points:

- Can be very inefficient
- Accurate nonparametric estimation of V needs high replication (small n_i, large m)

It is therefore better to incorporate correlation into the estimation of regression models. One should distinguish two cases:

– the correlation structure is known and then one uses weighted least squares to gain efficiency or

– it is unknown, but somehow structured, and then one uses MLE or RMLE.

Weighted Least Squares Estimation

Weighted least squares estimate of $\boldsymbol{\beta}$ minimizes

$$S(\boldsymbol{\beta}) = (\mathbf{y} - X\boldsymbol{\beta})'W(\mathbf{y} - X\boldsymbol{\beta}),$$

where W is a symmetric *weight matrix*.

Solution is

$$\hat{\boldsymbol{\beta}}_W = (X'WX)^{-1}X'W\mathbf{y}.$$

- unbiased: $\mathbb{E}(\hat{\boldsymbol{\beta}}_W) = \boldsymbol{\beta}$, for any choice of W,
- $\text{Var}(\hat{\boldsymbol{\beta}}_W) = \{(X'WX)^{-1}X'W\}V\{WX(X'WX)^{-1}\}$

Special cases

$W = I$**:** Ordinary least squares

- $\hat{\boldsymbol{\beta}}_W = \hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{Y}$,
- $\text{Var}(\hat{\boldsymbol{\beta}}_W) = (X'X)^{-1}X'VX(X'X)^{-1}$

$W = V^{-1}$**:** MLE under Gaussian assumptions with known V

- $\hat{\boldsymbol{\beta}}_W = (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y}$,
- $\text{Var}(\hat{\boldsymbol{\beta}}_W) = (X'V^{-1}X)^{-1}$.

When V is unknown, one usually assumes that the covariance structure of the sequence of measurements is specified by the values of unknown parameters $\boldsymbol{\alpha}$ which are estimated by appropriate methods (usually MLE or RMLE). One consider therefore

$$\mathbb{E}(\mathbf{Y}) = X\boldsymbol{\beta} \text{ and } \text{Var}(\mathbf{Y}) = V(\boldsymbol{\alpha})$$

Parametric models for covariance matrices

We have already discussed the *compound symmetry* model

$$Y_{ij} - \mu_{ij} = U_i + \epsilon_{ij}, \quad U_i \sim N(0, \nu^2),\ \epsilon_{ij} \sim N(0, \sigma^2),$$

used to model the pig weight data.

A common extension is the *random intercept and slope* model:

$$Y_{ij} - \mu_{ij} = U_i + V_i t_{ij} + \epsilon_{ij}, \quad \begin{bmatrix} U_i \\ V_i \end{bmatrix} \sim i.i.d. BVN(0, \Sigma),\ \epsilon_{ij} \sim N(0, \sigma^2),$$

which often fits short sequences well.

A general model for longitudinal data

A general model proposed by Diggle (1988) is

$$Y_{ij} - \mu_{ij} = U_i + W_i(t_{ij}) + \epsilon_{ij},$$

where

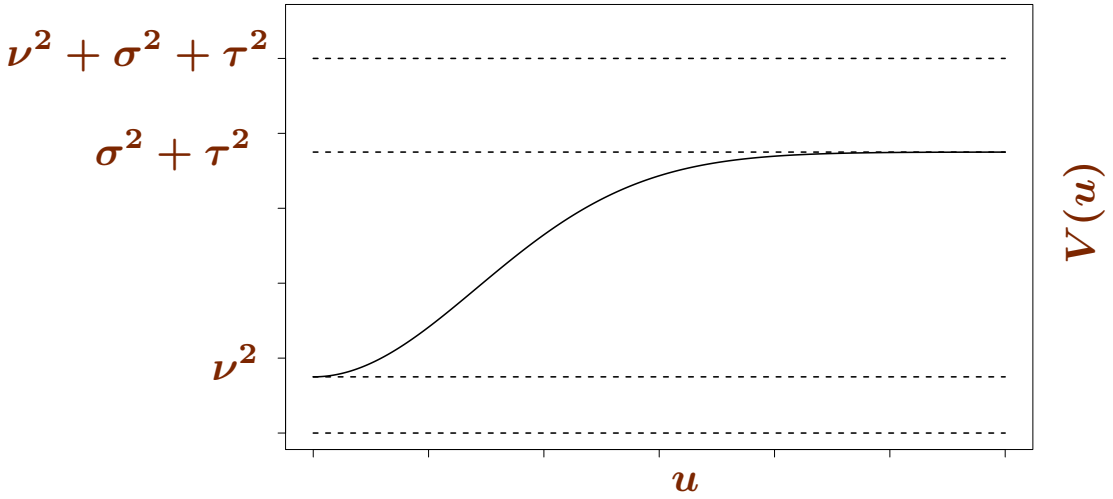
- random intercept $\mathbf{U}_i \sim i.i.d. N(0, \nu^2)$
- $W_i(t)$ i.i.d. centered continuous-time Gaussian processes modeling serial correlation, and
- measurement error $\epsilon_{ij} \sim N(0, \sigma^2)$.

If we further assume that the processes W_i are stationary then one can use the variogram to characterize these variance components.

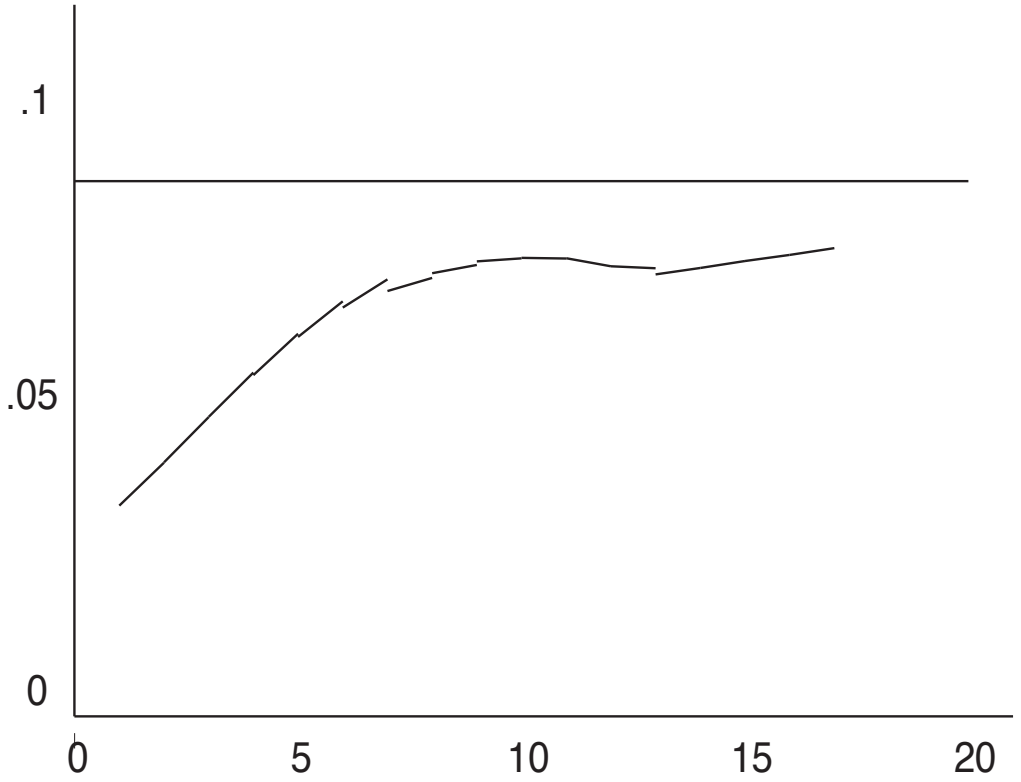
The variogram of the general model

$$Y_{ij} - \mu_{ij} = U_i + W_i(t_{ij}) + \epsilon_{ij}$$

$$V(u) = \sigma^2 + \tau^2(1 - \rho(u)), \quad \text{Var}(Y_{ij}) = \nu^2 + \tau^2 + \sigma^2.$$



Example on the protein content of milk



The overall variance of the residuals is 0.087; the variogram increases with lag (evidence of serial correlation); there is a random intercept (the variogram does not start at 0); evidence of measurement error (the variogram does not reach the total variance).

Estimation and Inference in Linear Mixed Effects Models

Mixed model methodology is one of the main contemporary tools for the analysis of longitudinal data. We have already illustrated this with the random intercept modeling of the pig weight data.

Just like the simple linear model,we can generalize mixed models to arbitrary design matrices. The covariance structure of the random effects vector can also be general. The resulting general linear mixed model is

$$\mathbf{Y} = X\boldsymbol{\beta} + Z\mathbf{U} + \boldsymbol{\epsilon},$$

where

$$\mathbb{E}\begin{bmatrix}\mathbf{U}\\ \boldsymbol{\epsilon}\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix} \text{ and } \text{Cov}\begin{bmatrix}\mathbf{U}\\ \boldsymbol{\epsilon}\end{bmatrix} = \begin{bmatrix}G & 0\\ 0 & R\end{bmatrix}$$

It is easy to see that the random intercept or the random intercept and slope models are special cases.

Estimation of fixed effects

One way to derive an estimate of $\boldsymbol{\beta}$ in the general linear mixed effects model (GLME) $\mathbf{Y} = X\boldsymbol{\beta} + Z\mathbf{U} + \boldsymbol{\epsilon}$ is to rewrite it as

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \quad \text{with } \boldsymbol{\epsilon}^* = Z\mathbf{U} + \boldsymbol{\epsilon}.$$

This is just a linear model with correlated errors, since

$$\text{Cov}(\boldsymbol{\epsilon}^*) = V = ZGZ' + R.$$

For a given V, the *Generalized least squares estimator* of $\boldsymbol{\beta}$ is nothing else than

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{V^{-1}} = (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{Y}$$

When the data is multivariate normal, the above estimator is the ML estimator of $\boldsymbol{\beta}$.

Best linear unbiased prediction

How about the *random effects* counterpart $\mathbf{U}$ in the GLME model? Maximum likelihood or least squares estimation is not defined for random effects and we need to appeal to *best prediction theory* :

$$\hat{\mathbf{U}} = \text{ best predictor of } \mathbf{U} \text{ given data } \mathbf{y},$$

i.e.

$$\hat{\mathbf{U}} = \mathbb{E}(\mathbf{U}|\mathbf{Y}).$$

It can be shown that

$$\text{BLUP}(\mathbf{U}) = \hat{\mathbf{U}} = GZ'V^{-1}(\mathbf{Y} - X\tilde{\boldsymbol{\beta}}).$$

However, most often in practice the covariance matrix V depends on unknown parameters $\boldsymbol{\alpha}$ which must be estimated.

Estimation of covariance matrices

Maximum likelihood and restricted maximum likelihood are the most common strategies for estimating the parameters in covariance matrices.

The MLE of V is based on the model $\mathbf{Y} \sim N(X\boldsymbol{\beta}, V)$. The log-likelihood of $\mathbf{y}$ under this model is

$$\mathcal{L}(\boldsymbol{\beta}, V) = \frac{1}{2}\left\{n\log(2\pi) + \log|V| + (\mathbf{y} - X\boldsymbol{\beta})'V^{-1}(\mathbf{y} - X\boldsymbol{\beta})\right\}$$

Optimizing in $\boldsymbol{\beta}$ for any fixed V, $\mathcal{L}(\boldsymbol{\beta}, V)$ is maximized over $\boldsymbol{\beta}$ by

$$\tilde{\boldsymbol{\beta}} = (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y}$$

On substitution into the log-likelihood we therefore obtain the *profile log-likelihood* for V:

$$\mathcal{L}_P(V) = -\frac{1}{2}\left\{\log|V| + (\mathbf{y} - X\tilde{\boldsymbol{\beta}})'V^{-1}(\mathbf{y} - X\tilde{\boldsymbol{\beta}})\right\} - n\log(2\pi)/2$$

ML estimates of the parameters in can be found by maximizing this expression over those parameters.

Restricted Maximum Likelihood

The dimensionality of optimization required to apply the ML principle is generally large. Moreover, simulations show that, generally, ML estimates of the variance components tend to be largely biased.

It is therefore advisable to "concentrate" the likelihood on the estimation of these components (and not on β) and this is achieved by using a *restricted maximum likelihood*. The resulting criterion function is the restricted log-likelihood

$$\mathcal{L}_R(V) = \mathcal{L}_P(V) - \frac{1}{2} \log |X'V^{-1}X|$$

Estimated BLUP (EBLUP)

The BLUP of $\mathbf{u}$ and the estimator $\tilde{\boldsymbol{\beta}}$ depend on G and R through

$$V = ZGZ' + R$$

whose parameters are typically estimated via ML or REML.

In practice the The BLUP of $\mathbf{u}$ and the estimator of $\boldsymbol{\beta}$ are replaced by

$$\hat{\boldsymbol{\beta}} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\mathbf{Y}$$

$$\hat{\mathbf{u}} = \hat{G}Z'\hat{V}^{-1}(\mathbf{Y} - X\hat{\boldsymbol{\beta}})$$

which are referred as *estimated BLUPs* estimates.

Standard Errors Estimation

For ordinary regression models the standard result is

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma_\epsilon^2 (X'X)^{-1}$$

so

$$\widehat{\text{st. dev}}(\hat{\beta}_k) = \sqrt{\hat{\sigma}_\epsilon^2} \sqrt{(X'X)^{-1}_{kk}}$$

For longitudinal models, $\text{Cov}(\hat{\boldsymbol{\beta}}) = (X'V{-}1X)^{-1}$, so

$$\widehat{\text{st. dev}}(\hat{\beta}_k) = \sqrt{\hat{\sigma}_\epsilon^2} \sqrt{(X'\hat{V}^{-1}X)^{-1}_{kk}}$$

These are the expressions used by the computer to provide confidence intervals and p-values.

Likelihood ratio tests

Estimation for Normal response longitudinal models is based on (restricted) maximum likelihood. Therefore the *likelihood ratio* paradigm for nested models may be used for hypothesis testing.

When testing H_0 : (smaller model) versus H_1 : *larger model*, classical theory says that under H_0 the test statistic

$$\lambda = -2\{\text{max. log- lik. under } H_0 - \text{max. log- lik. under } H_1 \text{ model}\} \sim \chi^2_k$$

where k is the difference in number of parameters between H_0 and H_1.

The assumptions of the classical theory don't hold in the longitudinal word due to lack of independence and parameters lying on boundaries of parameter spaces. Current ongoing research is confronting these issues.

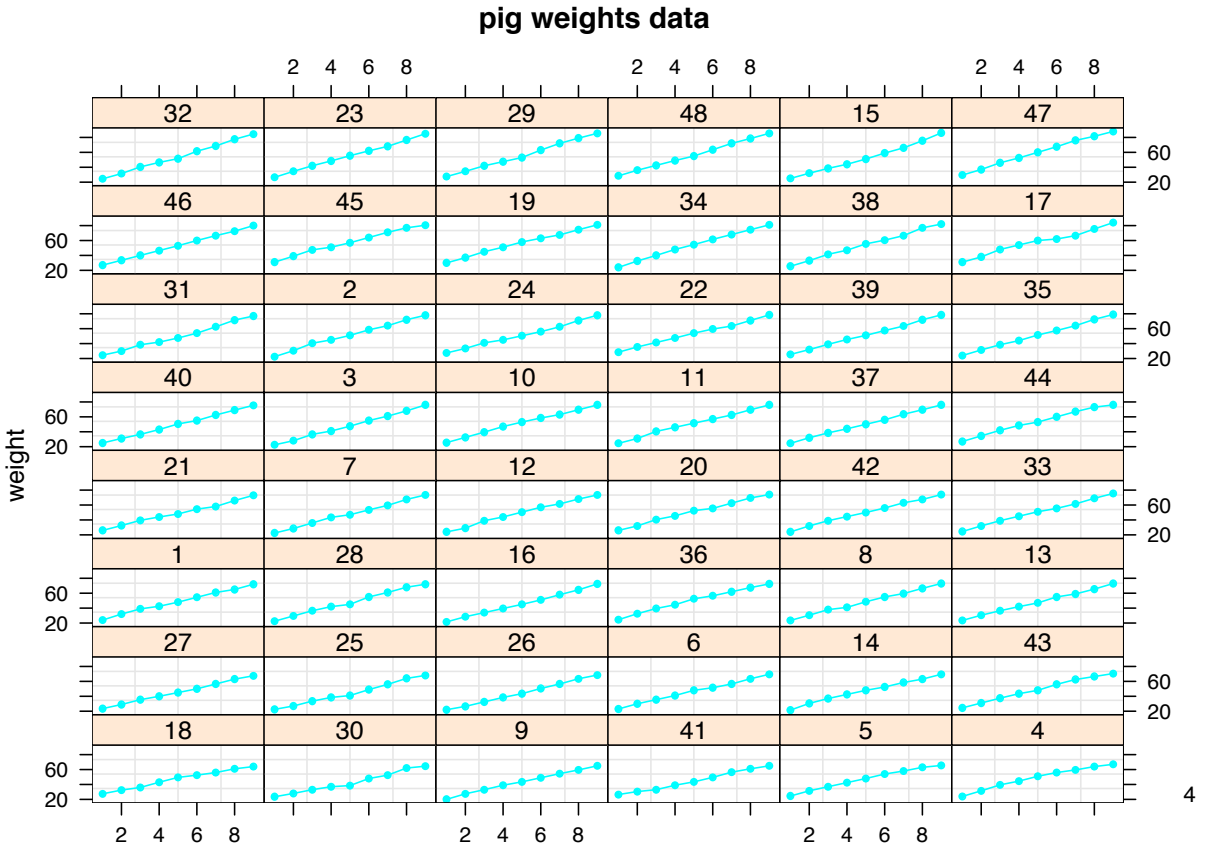
Computing

SInce the early 1990s software packages have started to support LDA. Examples are

- the `MIXED` and `NLMIXED` procedures in SAS
- the `lme( )` and `nlme( )` functions in R and S-plus
- the Stata package
- the Genstat package

Back to the Pig Weight Data

Recall the longitudinal data example involving 9 repeated measurements on 48 pigs.



A treillis plot of the data.

Random Intercept Model

We model the data by a random intercept model a special case of a mixed model using

$$\mathbf{Y} = \begin{bmatrix} Y_{1,1} \\ \vdots \\ Y_{1,9} \\ \vdots \\ Y_{48,1} \\ vdots \\ Y_{48,9} \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_9 \\ \vdots & \vdots \\ 1 & x_1 \\ \vdots & \vdots \\ 1 & x_9 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$Z = \begin{bmatrix} \mathbf{1}_{9\times 1} & \mathbf{0}_{9\times 1} & \dots & \mathbf{0}_{9\times 1} \\ \mathbf{0}_{9\times 1} & \mathbf{1}_{9\times 1} & \dots & \mathbf{0}_{9\times 1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{9\times 1} & \mathbf{0}_{9\times 1} & \dots & \mathbf{1}_{9\times 1} \end{bmatrix}, \quad \mathbf{U} = \begin{bmatrix} U_1 \\ \vdots \\ U_{48} \end{bmatrix},$$

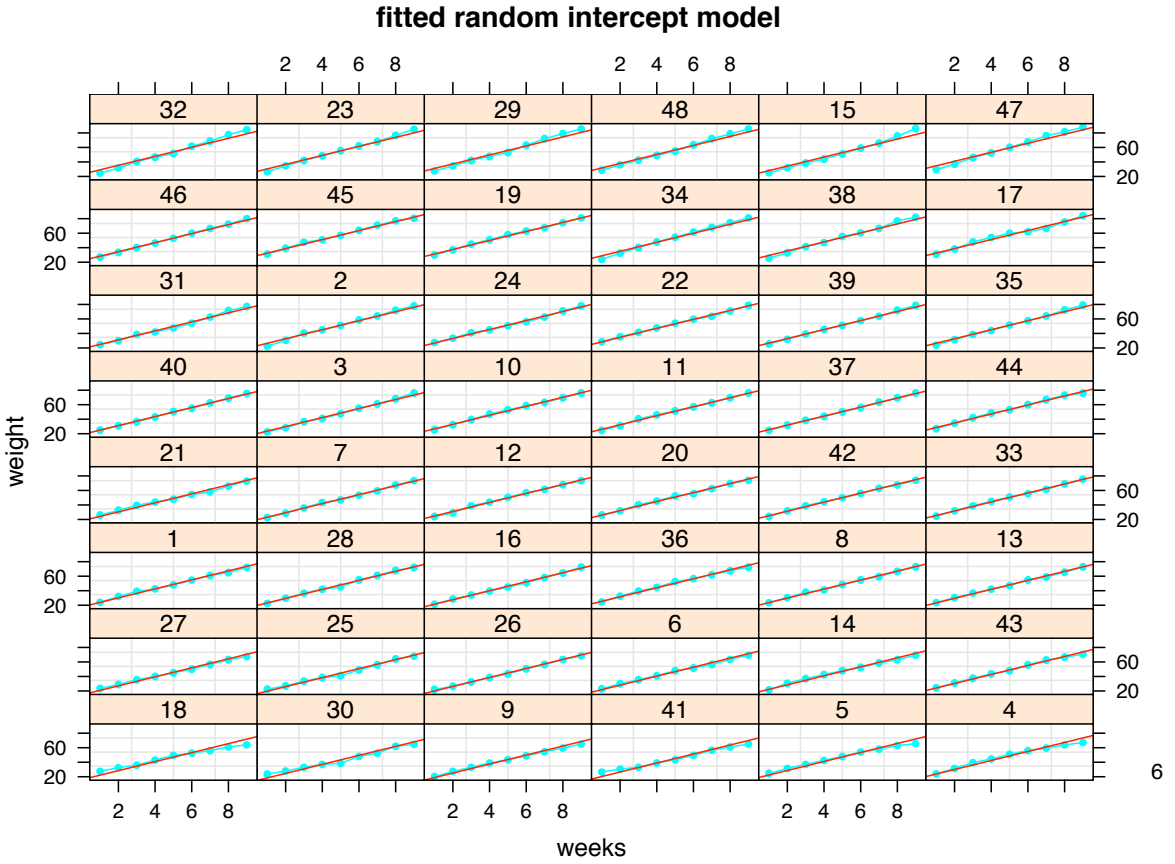
$$G = \sigma_U^2 \mathbf{I} \quad \text{and} \quad R = \sigma_\epsilon^2 \mathbf{I}$$

The estimates are

$$\hat{\beta}_0 = 19.36,\ \hat{\beta}_1 = 6.21,\ \hat{\sigma}_U^2 = 15.14 \text{ and } \hat{\sigma}_\epsilon^2 = 4.39,$$

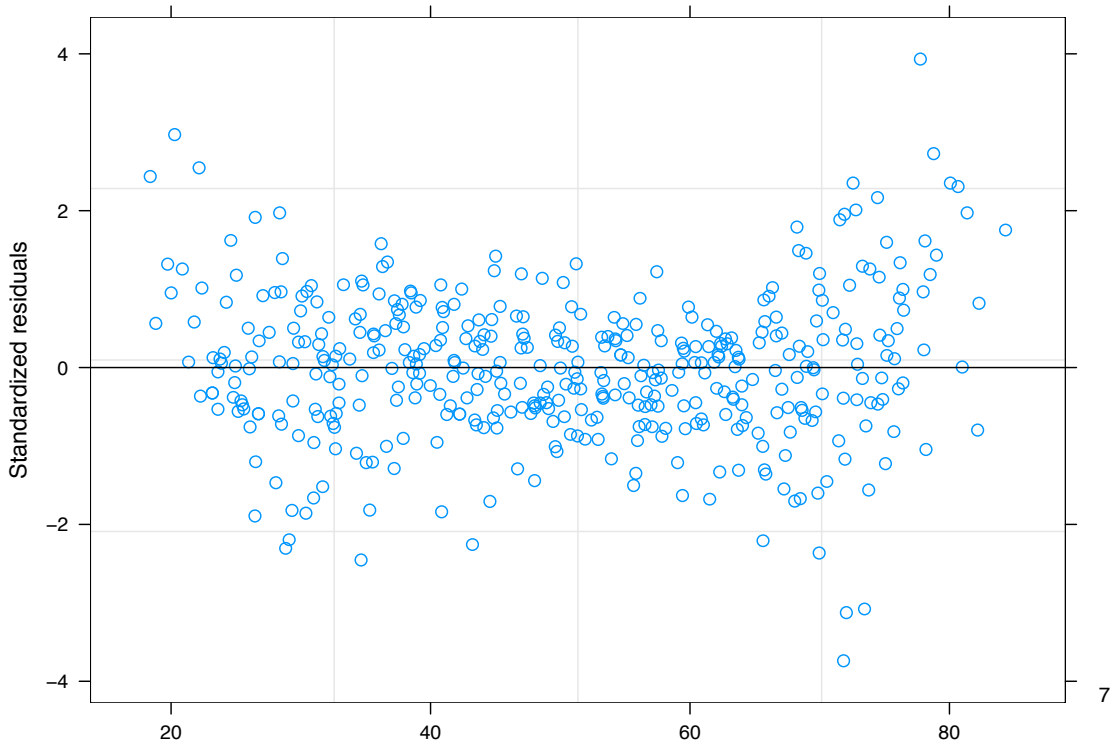
leading to $\hat{\rho} = 0.775$ and $\widehat{\text{st. dev}}(\hat{\beta}_1) = 0.0391$. The following two slides display the fitted random intercept model and the resulting residuals.

Fitted Random Intercept Model



Close inspection of the fitted lines shows that the parallel line assumption inherent in the random intercepts model is too restrictive.

Residuals in Fitted Random Intercept Model

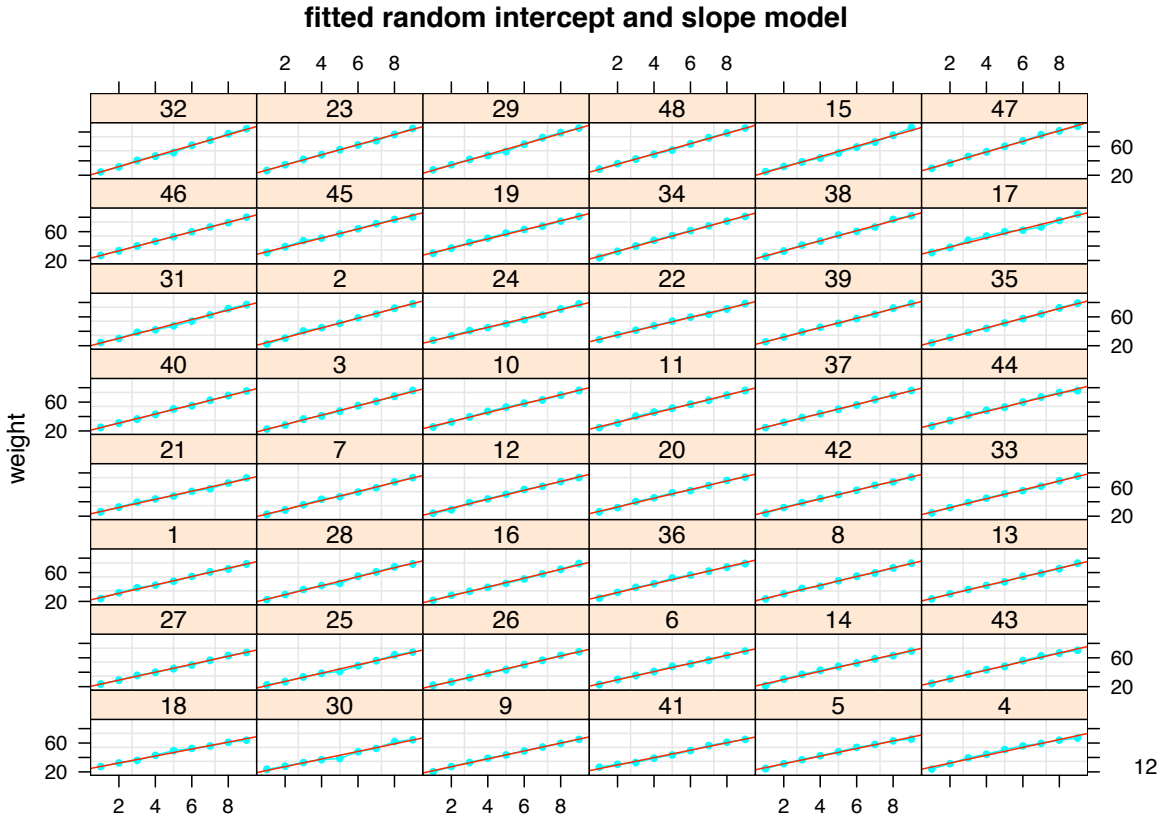


This is confirmed in the residual plot which shows a pronounced bow tie pattern.

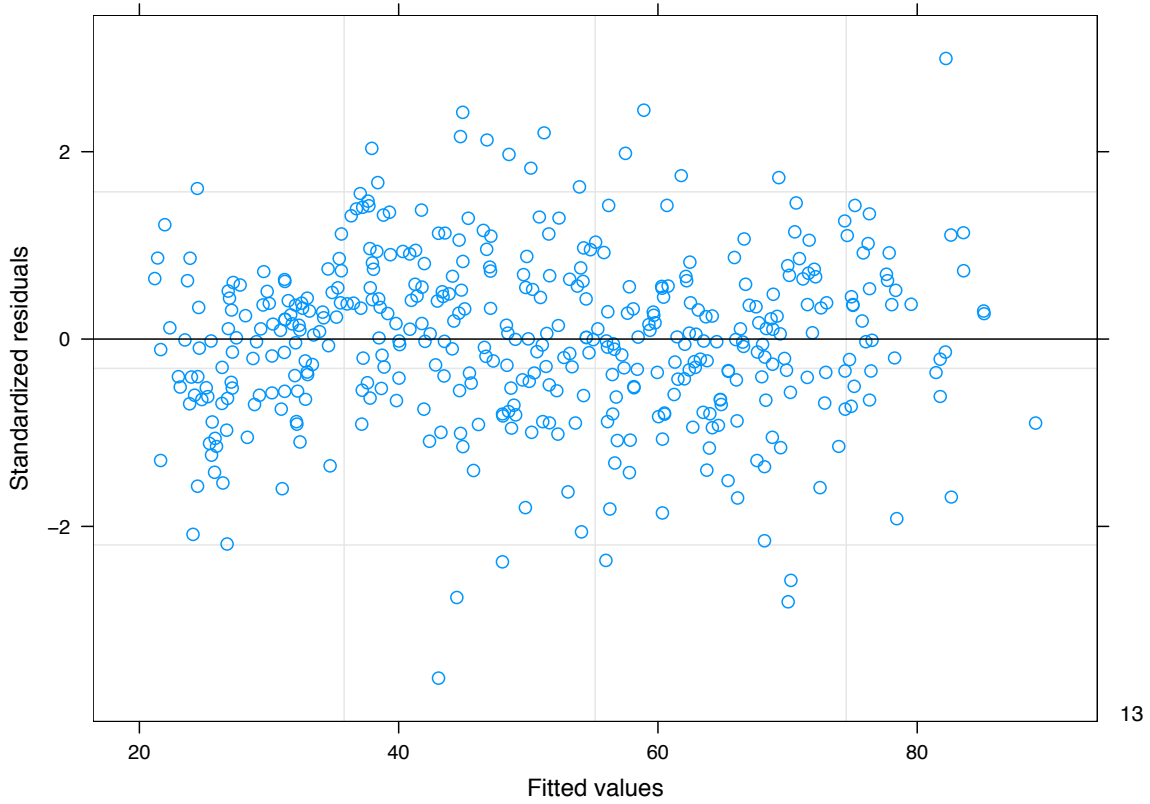
Random Intercept and Slopes Model

A remedy is to allow each pig to have his/her own slope,

$$Y_{ij} = U_i + \beta_0 + (\beta_1 + V_i)x_j + \epsilon_{ij}$$



Residuals in Fitted Random Intercept and Slope Model



The residual plot is now showing no systematic patterns. The second model is an improvement.

Other aspects (not seen in these lectures)

When dealing with non-Gaussian responses the classical generalized linear model (GLM) methodology unifies previously disparate methodologies for a wide range of problems, including:

- multiple regression/ANOVA (Gaussian responses)
- probit and logit regression (binary responses)
- log-linear modelling (categorical responses)
- Poisson regression (counted responses)
- survival analysis (non-negative continuous responses).

One can extend the classical GLM to analyse longitudinal data but we will not address this in these lectures as we will not also address any issues concerning missing values in longitudinal data.