EXPLORATORY / DESCRIPTIVE ANALYSIS OF LONGITUDINAL DATA

Characteristics of Longitudinal Data

- **Response variable:**
  - continuous: CD4+ count, weight of children
  - count: number of seizures in a two week period
  - ordinal: level of breastfeeding
  - binary: relief or no relief from pain
- **Time scale:** (what is "0-time"?)
  - age of subject
  - time since treatment start or since sero-conversion
- Number of **observations** per subject
  - fixed, e.g.:
    - 5 observations per Nepalese child
    - 4 follow-up observations per epileptic patient
  - variable, e.g., variable number of CD4+ counts due to differential follow-up, different numbers of visits, different times of sero-conversion

  - many, e.g., 19 weekly observations per cow: enough to model details of each cow's trajectory
  - few, e.g., 5 observations per Nepalese child: will capture only the simplest features of each child's trajectory (e.g., intercept and, maybe, slope)

- **Observation times** per subject
  - variable, e.g., ages of Nepalese children
  - fixed, e.g.:
    - once-every-two-weeks follow up of epileptic patients
    - weekly observations of milk protein data

- **Note:** Studies with equal numbers of observations and the same observation times for every subject are called **balanced**

Naive approaches to analyzing longitudinal data

- Ignore correlation
  – Biased SE, incorrect CI and p-value
- Using the end-points only
  – inefficient, not using complete data
- Derived variable approach
  – e.g. derive a slope for each individual, and obtain an average slope. Not taking into account the variability in estimating individual slopes
- Repeated ANOVA method
  – requires a balanced design, and very restrictive covariance structure

Notation for Longitudinal Data

- $i = 1, \ldots, m$ indexes **subjects**
- $j = 1, \ldots, n_i$ indexes **observations** for the ith subject
- $N = \sum_i^m n_i$ is the total number of observations
- t_{ij} is the actual **observation time** for the ith subject at the jth time
- Y_{ij} and y_{ij} are the random variable and observed **response** for the ith subject at the jth time
  - Upper-case Y to indicate random variable
  - Lower-case y to indicate its observed value
- $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$ or $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ is the $n_i \times 1$ **response vector** (of length n_i) for the ith subject
  - $\boldsymbol{Y}$ in bold to indicate vector.
- $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$ is the $p \times 1$ vector of covariates for the ith subject at the jth time
- $\mathcal{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})^T$ is the $n_i \times p$ design matrix of covariates for the ith subject

- The mean of and variance of $\boldsymbol{Y}_i$ given X_i are

$$\mathrm{E}(\boldsymbol{Y}_i|X_i) = \boldsymbol{\mu}_i \quad \text{and} \quad \mathrm{var}(\boldsymbol{Y}_i|X_i) = V_i \;\; (n_i \times n_i \;\; \text{matrix})$$

  where

$$\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^T$$

  is the $n_i \times 1$ **mean vector** for the ith subject

- This course will focus on regression models such as

$$\mathrm{E}(\boldsymbol{Y}_i|X_i) = \boldsymbol{\mu}_i = X_i\boldsymbol{\beta}$$

  $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the $p \times 1$ vector of coefficients

- In general:
  – capital letters for random variables or matrices
  – bold type for vectors

Software for Longitudinal Data Analysis

- SAS: PROC MIXED; PROC GENMOD; PROC NLMIXED; PROC GLIMMIX
- Stata:
  – first specify variables of subject ID and time using `xtset` *idvar timevar*
  – then you can use `xtdes`, `xtmixed`, `xtgee`, etc.
- R: `lme4`, `geepack` packages, etc.

Different software may have slightly different results due to computational reasons.

Data Format and Longitudinal Structure

We will generally work with data files of the following structure:

Subject	Observation	Time	Response	Covariates		
1	1	t_{11}	y_{11}	x_{111}	$\cdots$	x_{11p}
1	2	t_{12}	y_{12}	x_{121}	$\cdots$	x_{12p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	n_1	t_{1n_1}	y_{1n_1}	x_{1n_11}	$\cdots$	x_{1n_1p}
2	1	t_{21}	y_{21}	x_{211}	$\cdots$	x_{21p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2	n_2	t_{2n_2}	y_{2n_2}	x_{2n_21}	$\cdots$	x_{2n_2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	1	t_{m1}	y_{m1}	x_{m11}	$\cdots$	x_{m1p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	n_m	t_{mn_m}	y_{mn_m}	x_{mn_m1}	$\cdots$	x_{mn_mp}

- For example, here is part of the CD4+ data set:

	id	obs	timedays	cd4	sexpart	cesd
1.	10002	1	-271	548	10	8
2.	10002	2	-90	893	10	2
3.	10002	3	89	657	10	-1
4.	10005	1	-997	464	10	4
5.	10005	2	-822	845	10	-4
6.	10005	3	-81	752	10	-5
7.	10005	4	81	459	10	2
8.	10005	5	283	181	10	-3
9.	10005	6	459	434	10	-7
10.	10029	1	-453	846	10	18

	id	obs	timedays	cd4	sexpart	cesd
2370.	41829	9	1655	383	2	2
2371.	41844	1	-87	606	2	11
2372.	41844	2	87	570	5	10
2373.	41844	3	282	826	1	4
2374.	41844	4	562	983	1	8
2375.	41844	5	751	517	2	9
2376.	41844	6	1249	462	2	0

- Note: Baseline covariates (if any) are repeated for each observation, even though they are constant within subject
- The variable `obs` is optional
- Two important variables are `id` and `timedays` because they identify:
  – the **subject** to which the observation belongs
  – the **time** at which the observation occurred
- These data are in "long" format
- An alternative data structure is to have the repeated measures $y_{i1}, \ldots, y_{in_i}$ running across the rows as variables.
  – This is called "wide" format; we will not use this format much
  – Transforms the data between long and wide formats:
    * Stata: `reshape`
    * SAS: PROC TRANSPOSE
    * R: `reshape()`

Longitudinal Data Structure: Another Example

- **Example:** Protein content in cows' milk depends on diet
- Scientific goal: Relationship of time profile (since calving) of protein content of milk to diet (lupins, barley, mixed)
- The time scale of interest is number of weeks since calving, with preliminary inspection of cows data:
  – the study ran 19 calendar weeks
  – cows calving in the first week of the study have complete data
  – cows calving later have missing data

Frequency table for missing pattern:

Freq.	Percent	Cum.	Pattern
37	46.84	46.84	1111111111111111
18	22.78	69.62	11111111111111.....
8	10.13	79.75	111111111111111....
4	5.06	84.81	1111111111111111...
4	5.06	89.87	1111111111111111.
1	1.27	91.14	1.1111111111111111
1	1.27	92.41	1111.111111111.....
1	1.27	93.67	111111..1.11111....
1	1.27	94.94	111111.111111111111
1	1.27	96.20	1111111.11111111111
1	1.27	97.47	11111111.1.111.....
1	1.27	98.73	11111111.1111111111
1	1.27	100.00	111111111111.11111.
79	100.00		XXXXXXXXXXXXXXXXXX

- for each cow, "1" indicates observed at this week, "." indicates missing at this week
- Note: If calving date is related to protein content of milk, these missing data could introduce bias in estimates of protein content, especially at later weeks

EXPLORATORY ANALYSIS OF LONGITUDINAL DATA

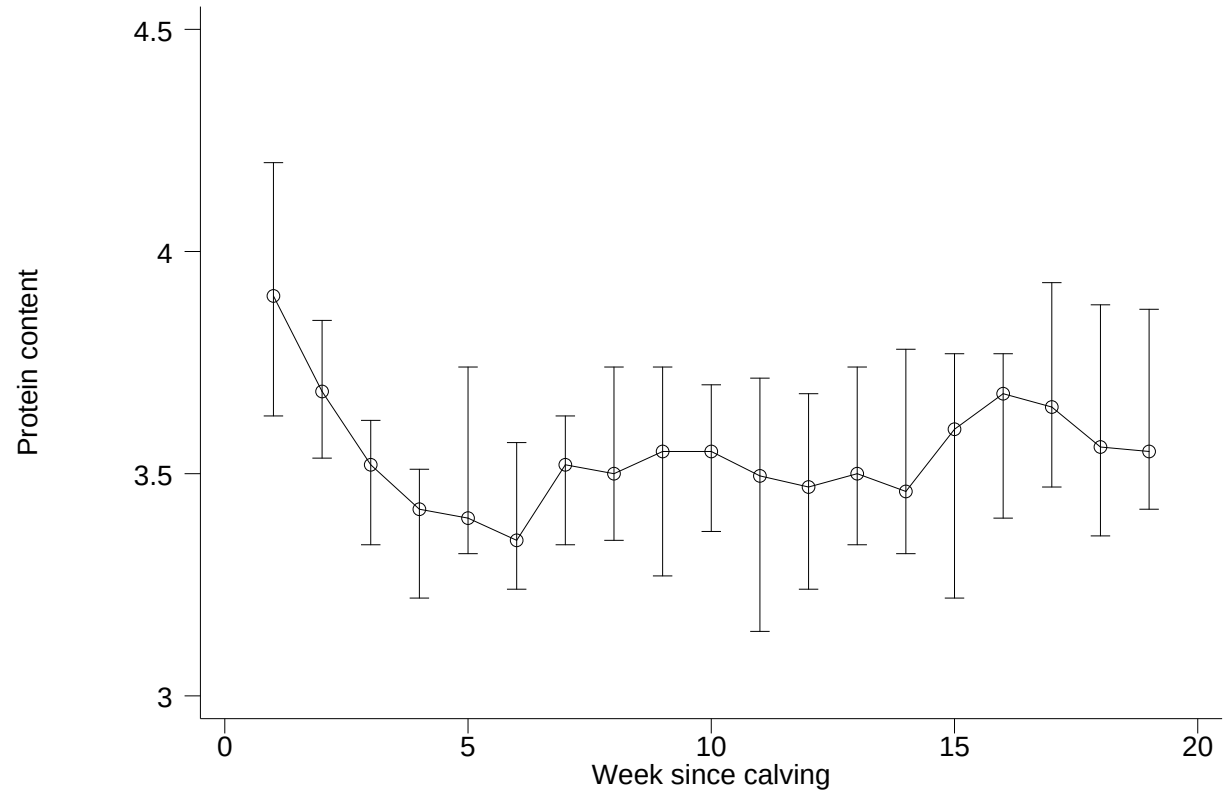
Population Average Patterns

Start exploring the data:

- Trends in data, ignoring some longitudinal structure for now (i.e., ignoring the fact that the data are repeated measures on a set of individuals):
  - Treat data as "cross-sectional"
  - Not a bad way to start exploring the data (because time trends are elucidated)
  - Present (features of) the **distribution** of response Y across time
    - mean or median
    - other quantiles or displays of dispersion
  - tools: box plots, "error-bar" plots, smooth model fits

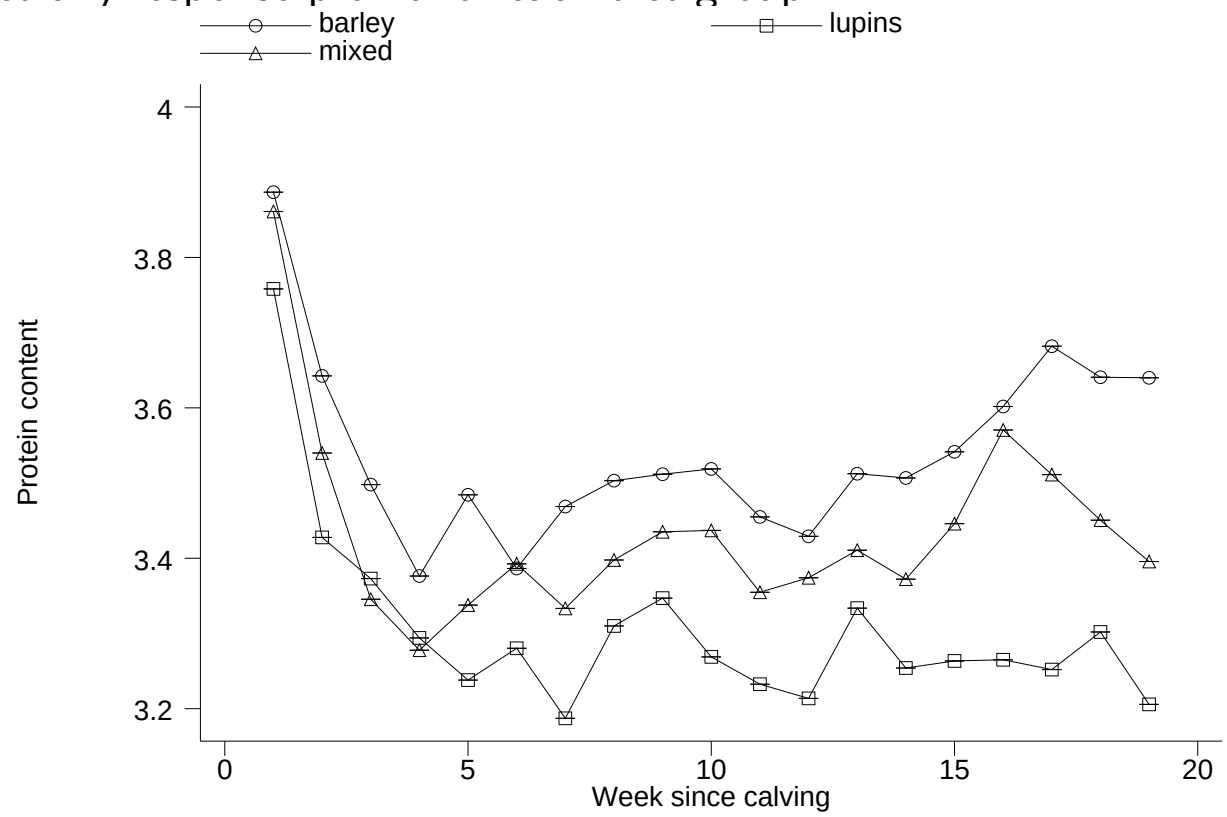
- **Example 1: "Error-bar" plots** for the protein content of cows on barley diet:

Median (interquartile range = 25% and 75% percentile) protein content by week since calving



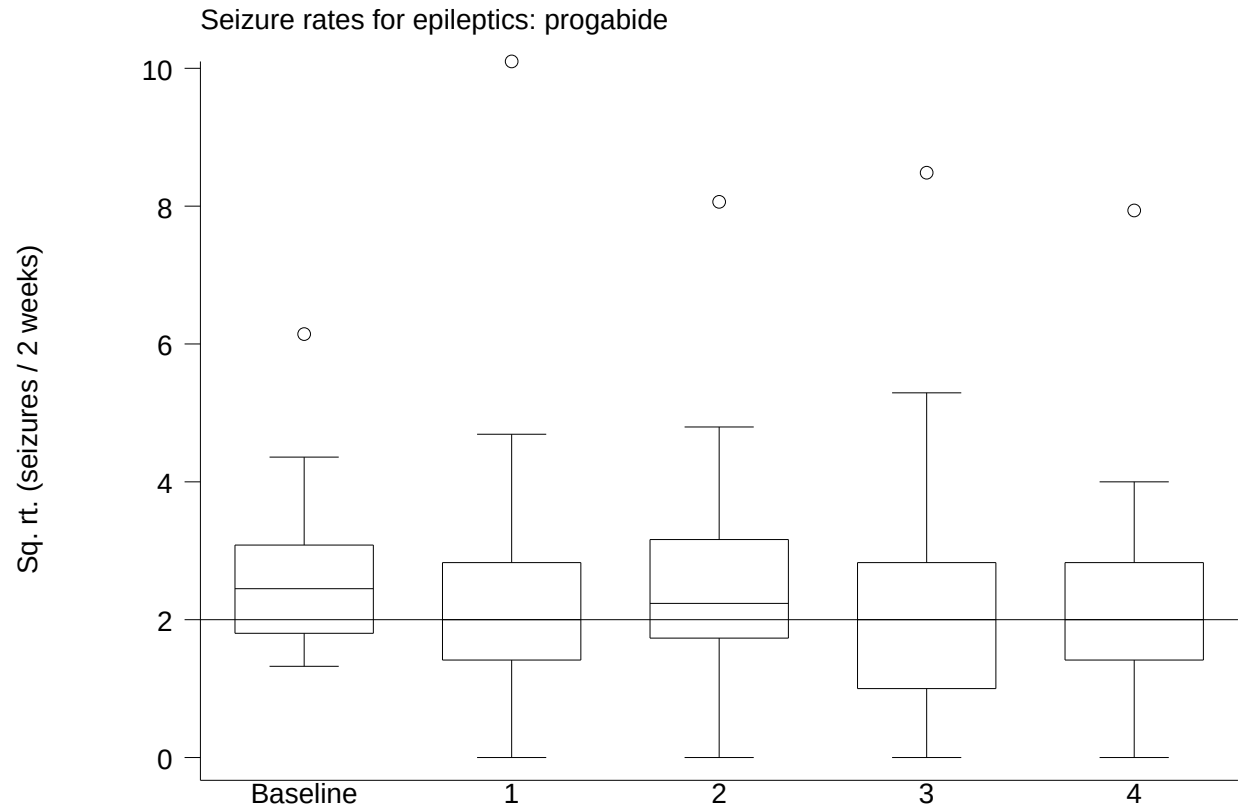
– We are presenting **distributions** here, so "error-bars" should plot **data features** such as standard deviation, interquartile range, etc.
  – **not** inferential indices such as standard errors or confidence intervals
– Outliers could be added to plot (with effort)

– As the goal of study is to compare three groups, we examine the *mean* (versus median) response profile for each diet group



– Preliminary observations:
  * Protein content drops strongly over first several weeks, then levels off
  * Cows on diets with more barley have higher protein content (on average)
  * Mean response **profiles** for each group follow a similar pattern over time

- **Example 2: Side-by-side box plots** for the seizure data for subjects on progabide:



  - Here $y =$ number of seizures in each time interval, and we plotted

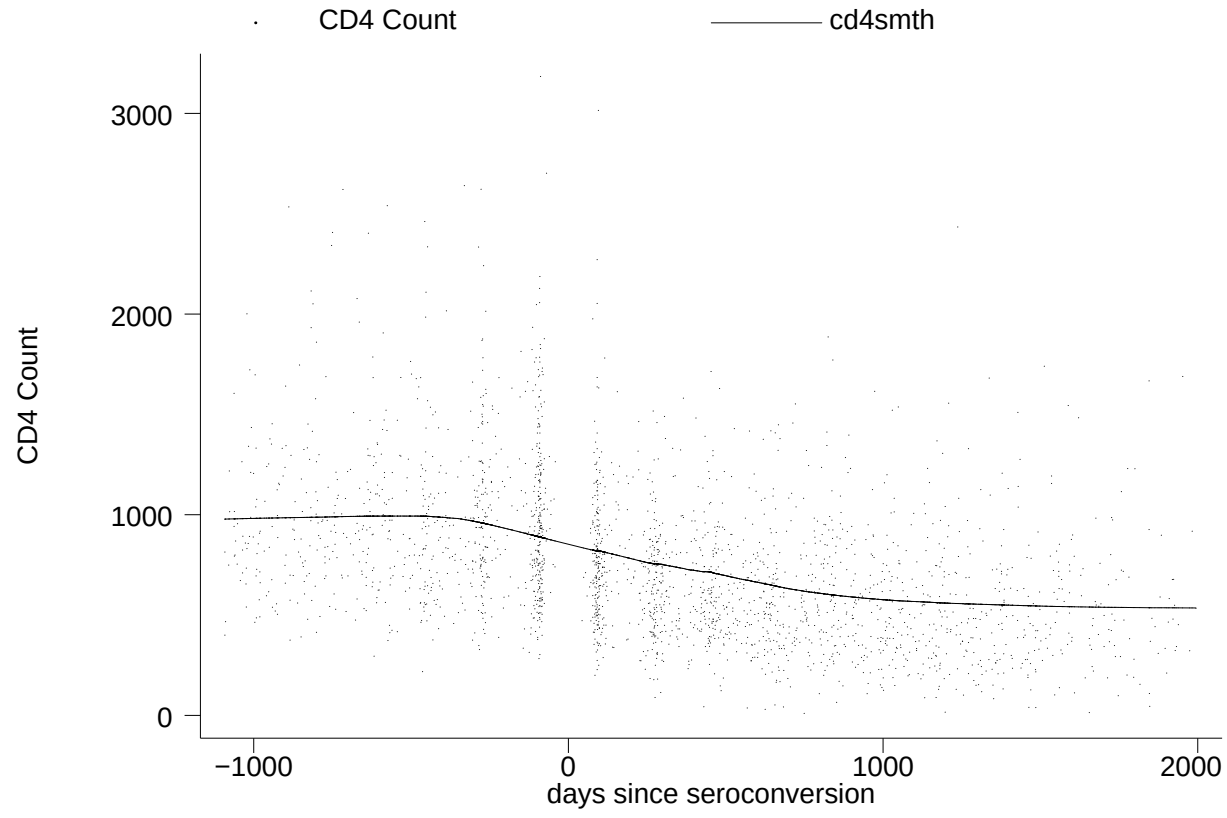
$$y^* = \sqrt{\frac{y}{\text{length of interval}}}$$

($\sqrt{}$ is variance-stabilizing transformation for Poisson data)

- Shows variation in central tendency (median), dispersion (inter-quartile range) and in outliers with respect to time
- Plot useful if # of time points small or can make time "groups"

- For the seizure and cows examples, time is fairly discrete; what if it is more continuous?
  - E.g., if each subject admits his/her own set of observation times?
  - Use **scatterplot smoothers**.

- **Example 3: Smooth model fit** for the mean CD4+ count of men before and after HIV seroconversion



  - This plot shows **all** of the data points ("micro") and also highlights an important **central feature** of the data ("macro"): the mean CD4+ count as a (smooth) function of time

– The mean CD4+ count is expressed as a flexible (i.e., not necessarily linear) function $\mu(\cdot)$ of time by the **non-parametric regression model**

$$\text{CD4} = \mu(\text{timedays}) + \epsilon$$

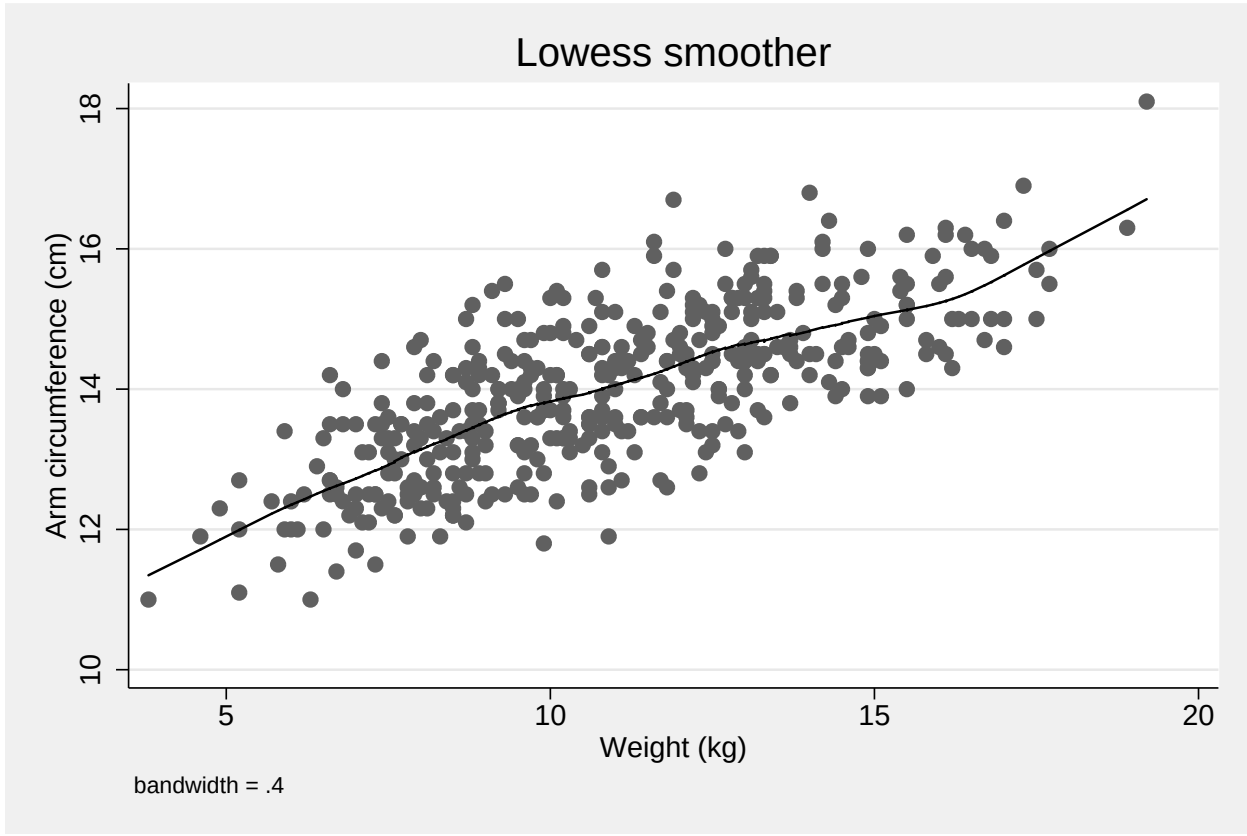
Scatterplot smoothers fit such models to data such as these
– Eg, use lowess (loess) function in R/Stata

– This type of plot is very useful when times are unequally-spaced

- General notions of exploratory LDA thus far:
  Questions to be answered at the outset of any analysis:
  - What is **nature** of response variable? Continuous? Count? Binary?
  - What is the pattern of **number** and **timing** of measurements in the **design**?
    - number and frequency?
    - same for all (**balanced**)?
    - mild **imbalance** or severe? missing data?
  - What is the distribution of responses by time:
    - box-plot or similar for highly balanced settings
    - scatter plot for unbalanced settings

    These are called **time plots** because they plot the data or some summary thereof versus time
  - What is the population average or mean trend with time?
    - add mean or median curves to to time plots

- So far, we have investigated the relationship of **response** to **time** ...
- What about relationships among variables other than time?
- **Example 4:** Arm circumference and weight in Nepalese girls
  - Arm circumference is sometimes used as a population screening tool for malnourishment.
    - It is much easier, and perhaps more accurate, to measure than weight
    - However, its utility is limited to the degree that it reflects weight
  - It is therefore of interest to examine the relationship of arm circumference to weight
  - A simple graphic display of this relationship is given by plotting `arm` versus `wt` with lowess (we will look at girls here):



– However, recall that the age distribution in these data varies from less than one to more than five years, and both arm circumference and weight are certainly related to age
– We can use a smoothing model fit to **remove** the age trends in arm circumference and in weight . . . essentially performing a partial residual analysis

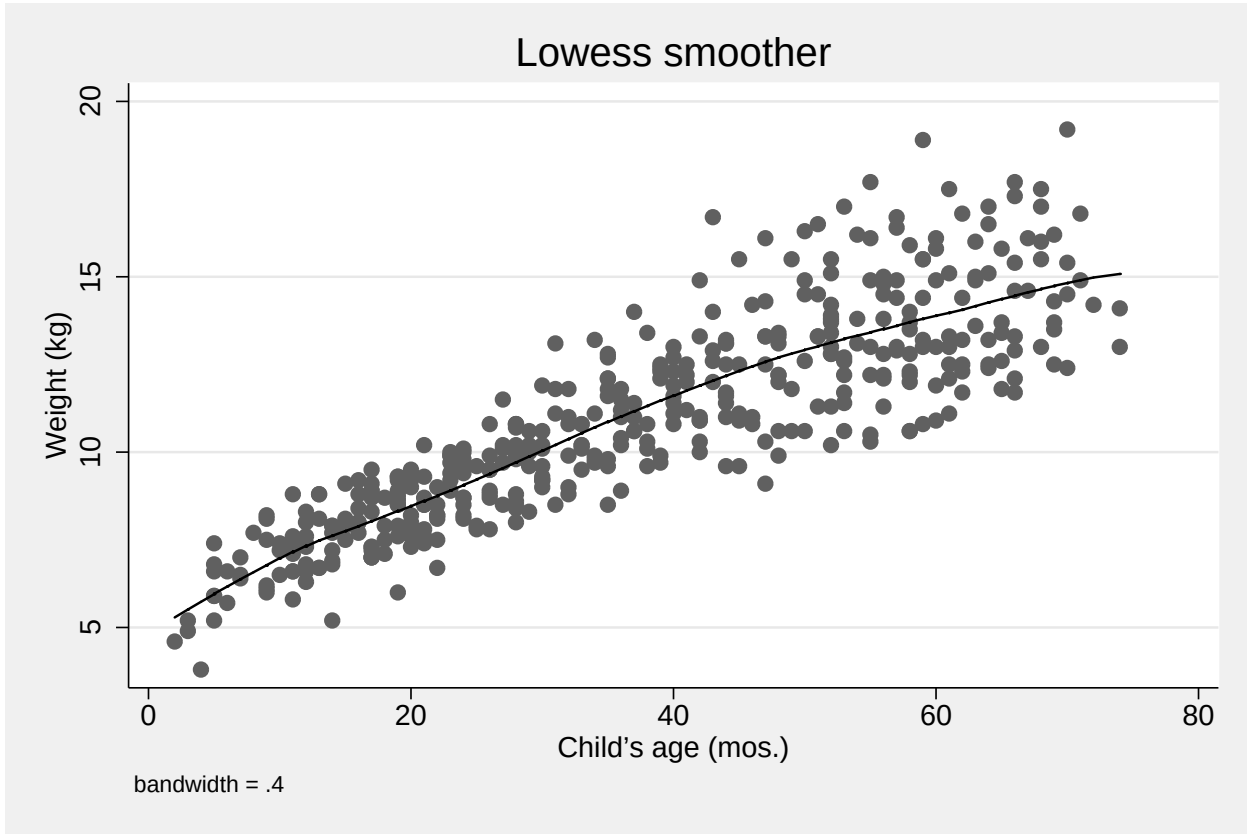
– Use flexible smoothing model here (instead of linear model) for exploration

∗ First fit the models

$$\text{arm} = \mu_1(\text{age}) + \epsilon_1$$

and

$$\text{wt} = \mu_2(\text{age}) + \epsilon_2$$



* Now compute the residuals for arm circumference and for weight

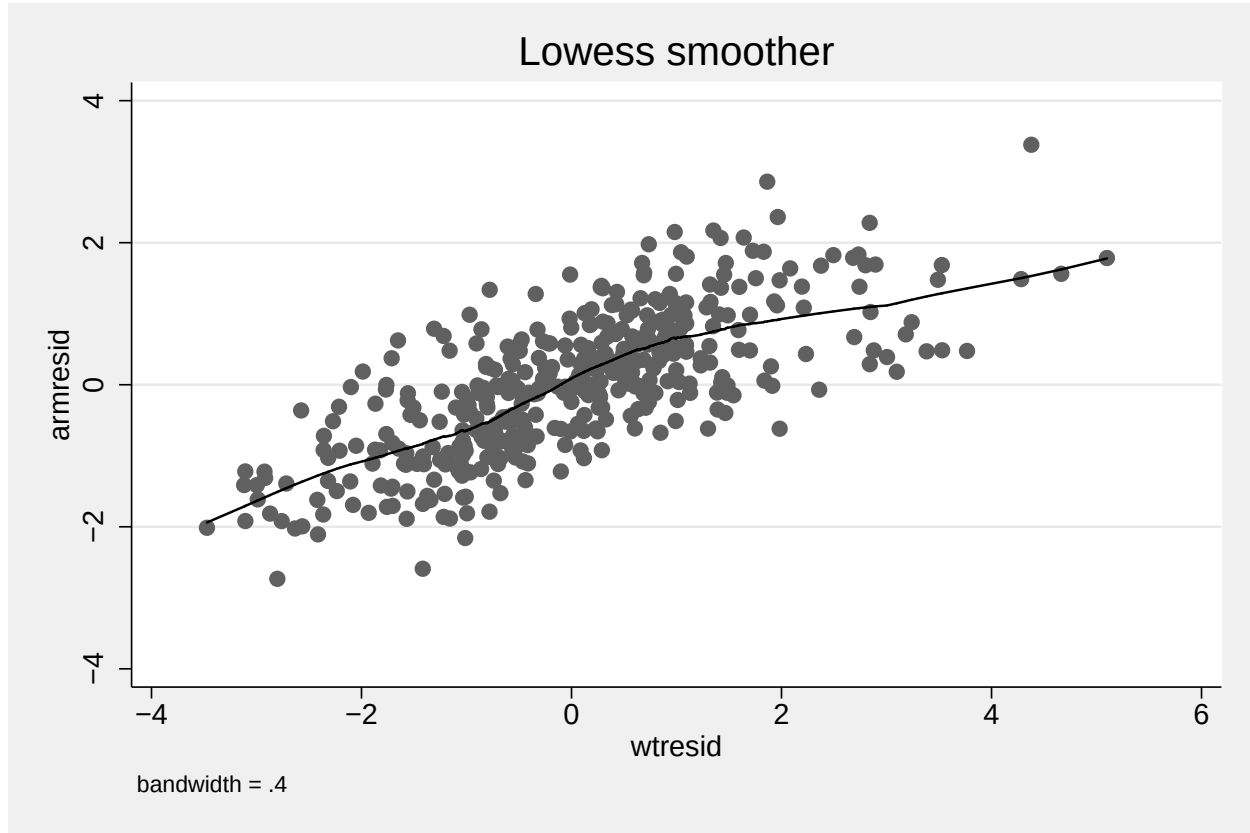
$$\text{resid}_{\text{arm}} = \text{arm} - \hat{\mu}_1(\text{age})$$

and

$$\text{resid}_{\text{wt}} = \text{wt} - \hat{\mu}_2(\text{age})$$

  * Finally, plot the arm circumference residuals versus the weight residuals

– The result is a plot of arm circumference versus weight, **adjusted for age**:



– In this case, the age-adjusted relationship between arm circumference to weight appears stronger than the unadjusted relationship, at least for the lower and middle weights

- All of these plots explore **cross-sectional** features of the data (i.e., they ignore the **longitudinal structure** of the data):
  - Here is why: Because the mean, median or any other feature of the data at each time point would be the same (on average) if one had collected cross-sectional data at each time point
  - E.g., the CD4+ plot above could be constructed with $n = 2376$ **independent** observations of CD4+ count and time since seroconversion
  - This is OK for exploratory purposes

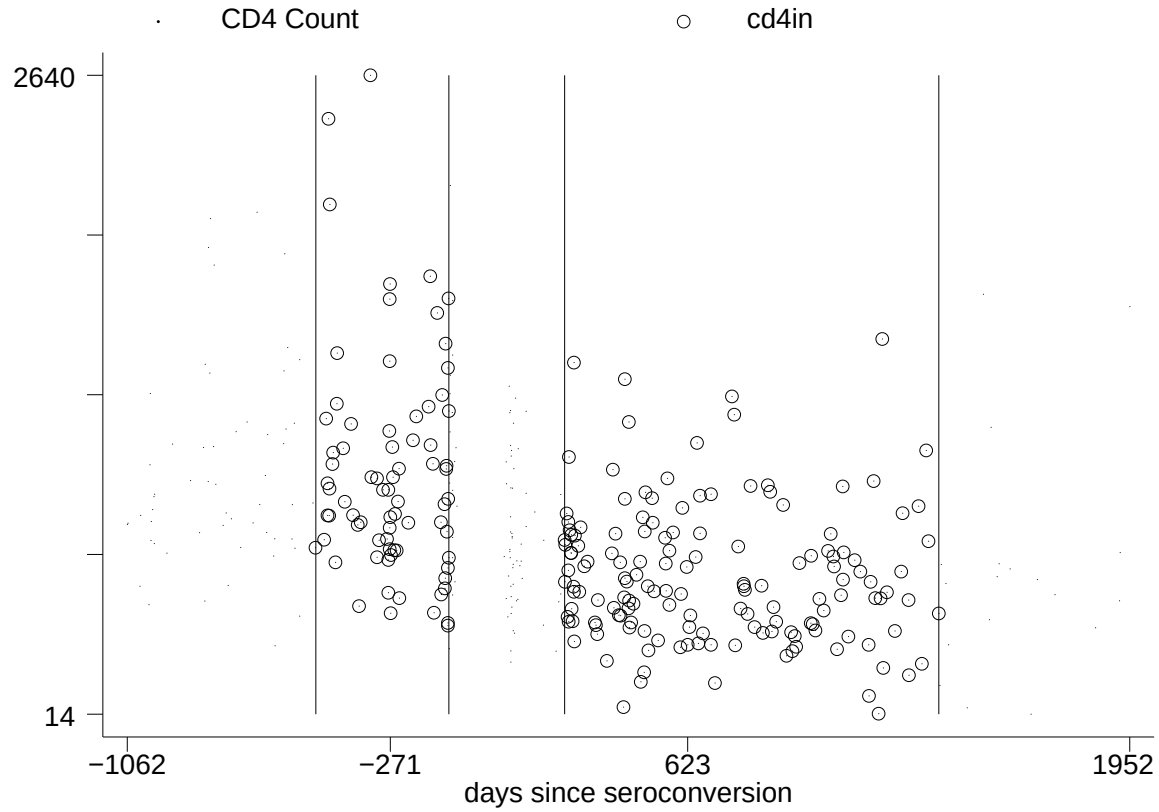
Supplement: How Smoothers Work

- Examine the CD4+ data with only one randomly-selected observation per subject
- We wish to fit a flexible (smooth) model of mean CD4+ (Y) to time t

$$Y = \mu(t) + \epsilon$$

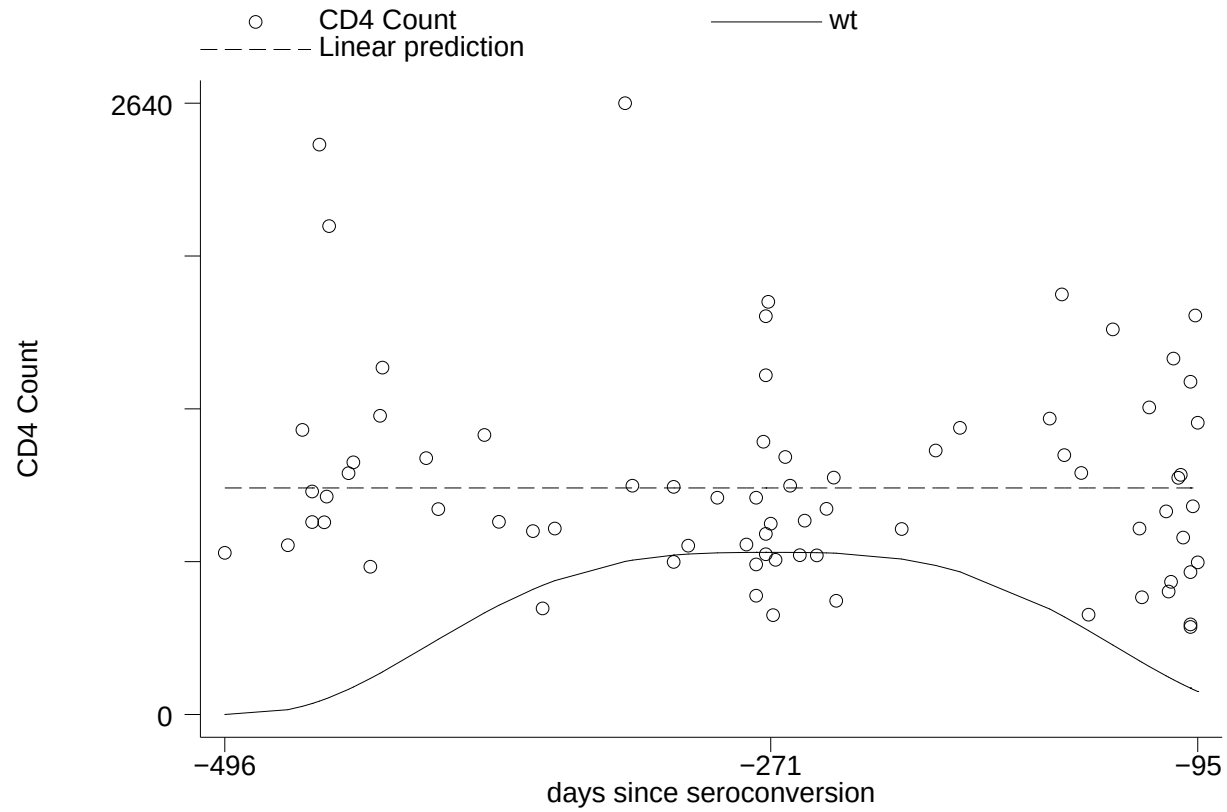
- This means estimating $\mu(t)$ for every value of t. For a given t^*, a **band** of observations is defined and these are used for estimation of $\mu(t^*)$ [i.e., $\mu(t)$ at t^*]

Here are two bands around $t^* = -271$ and $t^* = 623$ days:



The $t^* = -271$ band is a 20-percent band (the closest 20% of the data to the target time, t^*), while the $t^* = 623$ band is a 40 percent band

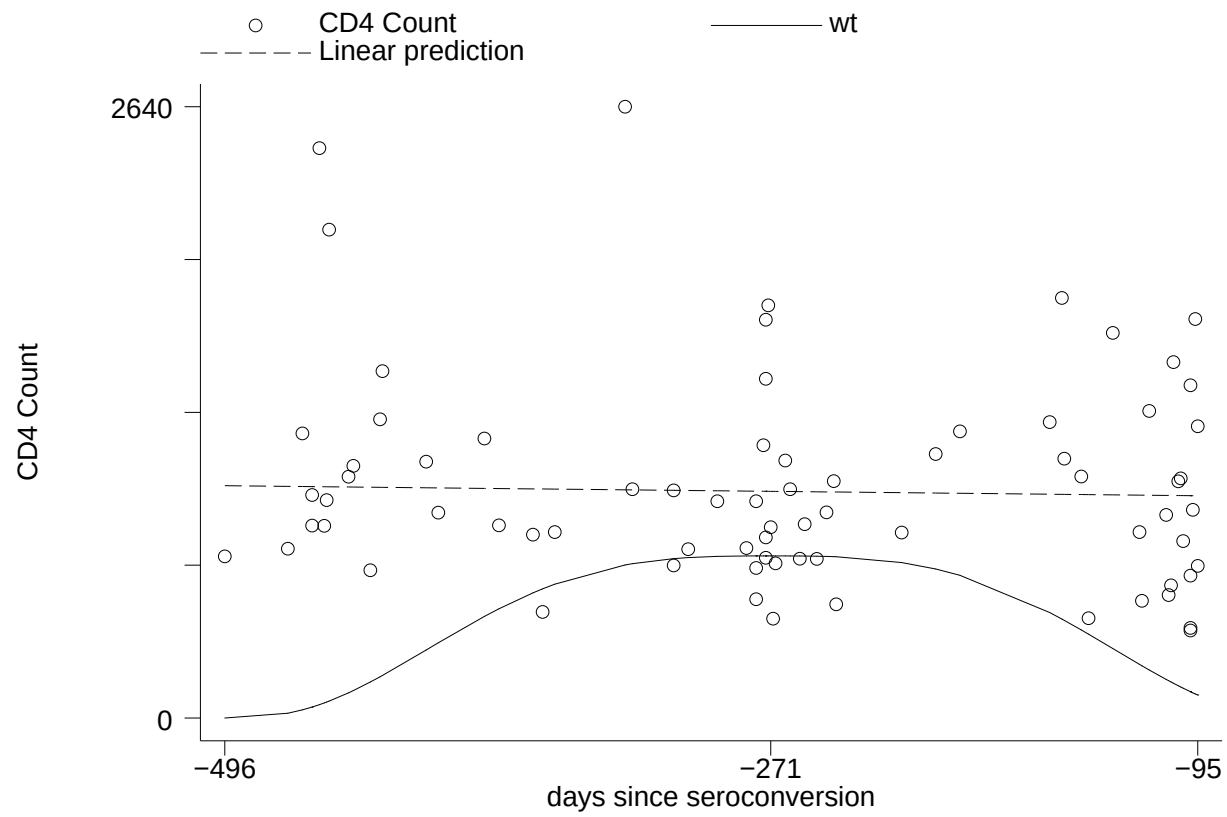
- Now just look at the band around $t^* = 271$:



- A **weighting function** (based on a **kernel**, eg, Gaussian kernel) is defined within that band that gives heavier weights to observations closer to $t^* = -271$
- Then, a **weighted mean** (dashed horizontal line above) is computed for the observations in the band

- This weighted mean is the estimate of $\mu(t^*)$
- This is **repeated** for every value of t in the data set, resulting in the fitted smooth fit $\hat{\mu}(t)$
- This is called **kernel regression**

- An improved version of kernel regression does not compute a weighted mean, but instead computes a **weighted OLS regression** for each t^*:



  The regression line at $t^* = -271$ is the estimate of $\mu(t)$ at t^* (the fitted line has a slight negative slope, but in this case it makes very little difference)

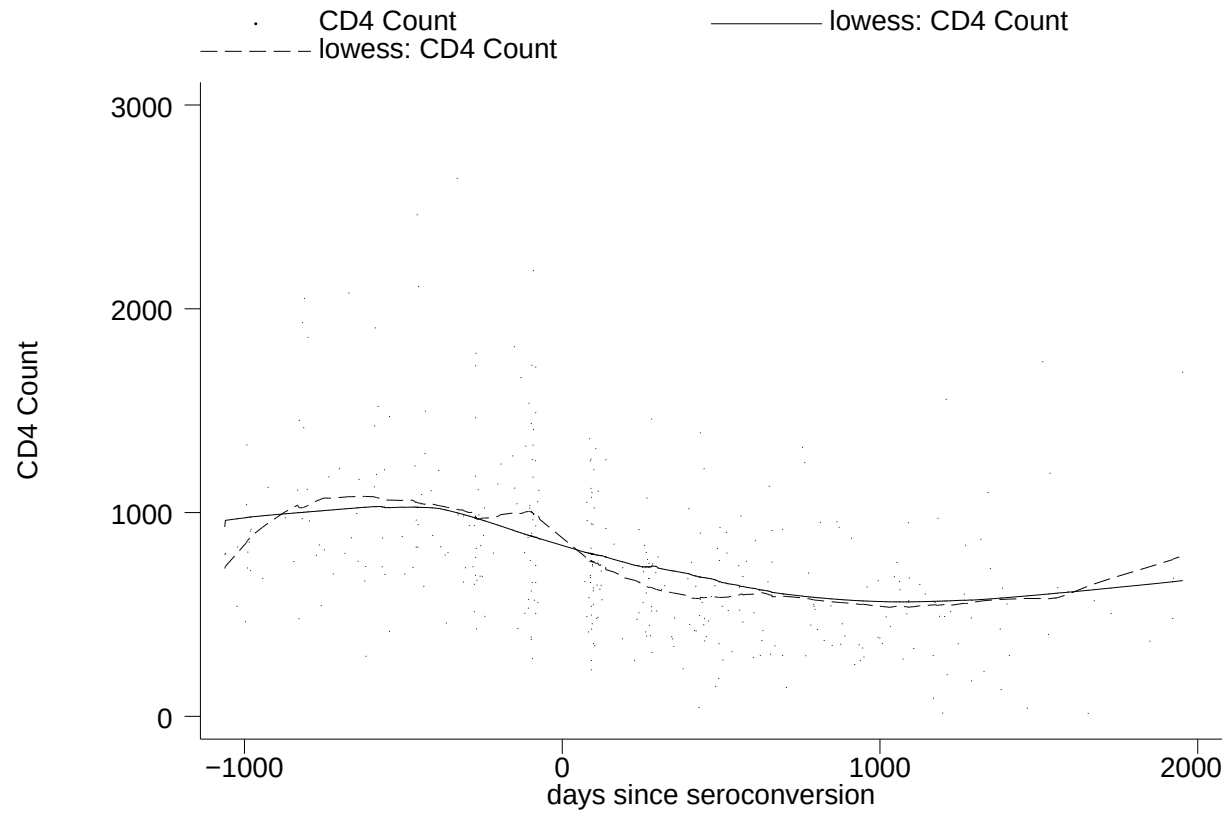
- This improvement is called **lowess** where "lowe" stands for "locally weighted"

- There are many omitted details in this discussion:
  - choice of weighting function (kernel)
  - how to define the bands
  - what to do at the edges of the data
  - reducing effects of outliers
  - choice of bandwidth

The final result is often pretty robust to these choices

- **Choosing the bandwidth**, however, is important and statistically tricky, but for exploratory analysis, we can do pretty well with our own eyes

  Here are plots with two smooths: bandwidth 80 percent and bandwidth 20 percent:

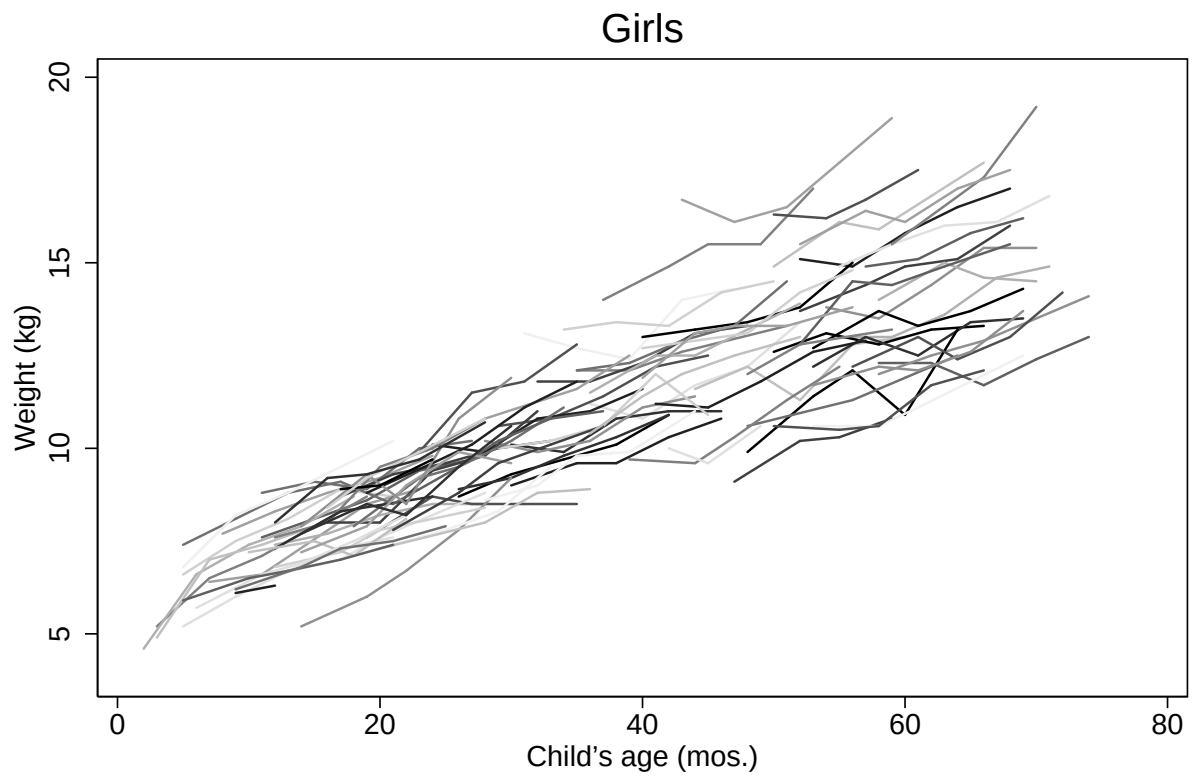


Bandwidth of 20 percent means that each band contains 20 percent of the data, resulting in less smoothing than bandwidth 80 percent

Adding Longitudinal Information to Plots

- Plots above present the data as if they were **cross-sectional**
- We would like to augment those plots with **longitudinal information**
- An obvious attempt at this was presented earlier: Plot each subjects trajectory over time (i.e., a **time plot** for each subject)

  Example: Weight of Nepalese girls



We can extract **some** information from this:

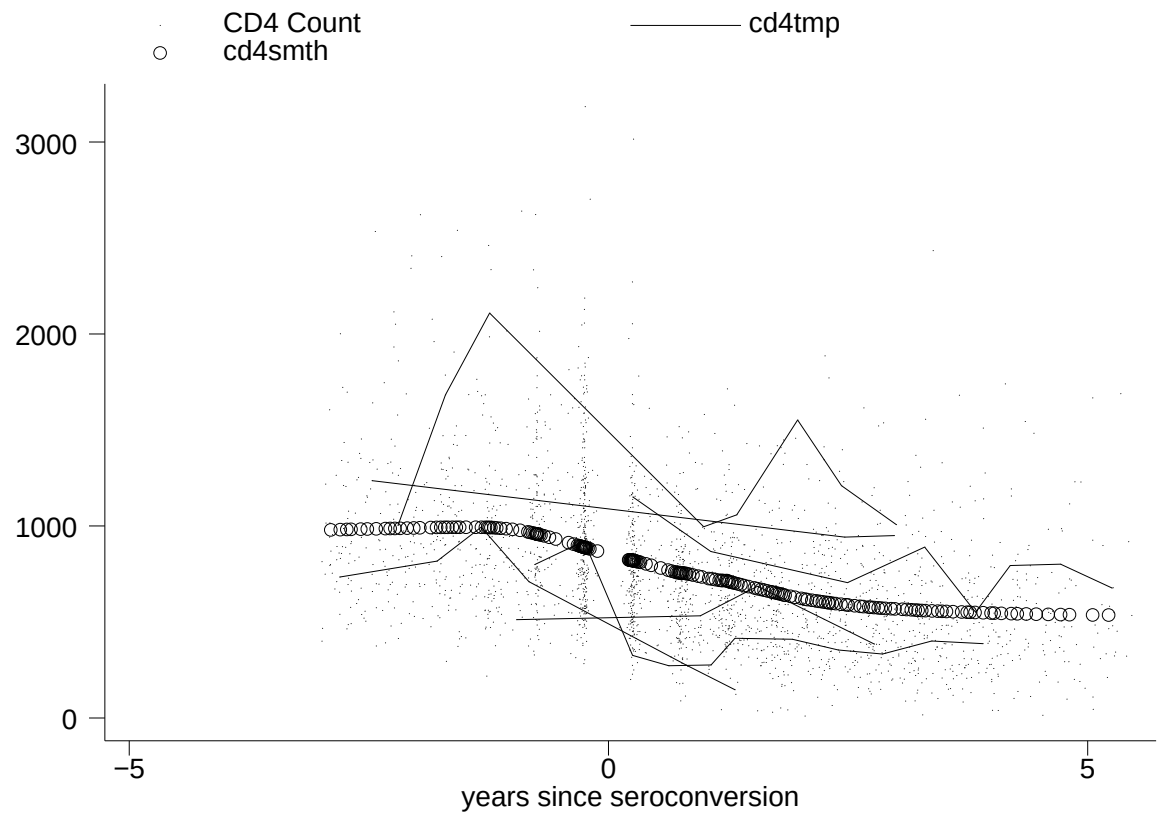
- subjects for the most part gain weight with age
- the variance in weights increases with age
- the data exhibit some "tracking"

In short, we can begin to see the dependence among repeated measures on the same individual, separating **within**- and **between**-subject variability

- Problem: The plot may be too busy

  Solution: Plot the trajectories (**line plots**) for only a **randomly-selected subset** of subjects

- **Example:** In the CD4+ data set, suppose we were interested in displaying the distribution of **responses to HIV seroconversion** in terms of subjects' trajectories of CD4+ counts


CD4 Count
cd4smth
cd4tmp
3000
2000
1000
0
−5
0
5
years since seroconversion

- Notes: We can now see that:
  1. there is considerable variability within subject in CD4+ count
  2. there is also considerable variability between subjects in CD4+ count
  3. there is variability in degree of response in CD4+ count to HIV infection
  4. overall drop in CD4+ count is observed for most subjects
  5. one subject has a very strange result and could be considered an outlier
- Caveat: While the time-plots of a small randomly-selected subset of subjects are likely to convey some longitudinal information, they are not likely to:
  – capture the full distribution of trajectories
  – identify unusual or outlying observations

  For this, it might be useful to chose subjects based on a **ranking of within-subject statistics** (interested students can see Diggle, Heagerty, Liang and Zeger)

EXPLORING LONGITUDINAL DATA

Separating Longitudinal and Cross-sectional Patterns

Within and Between Subject Variance

- **Example:** Weight of Nepalese girls

  We have already seen that a large part of the variance in weight is due to age differences among the subject and observation times

- What about the **residual variance**, removing the effects of age?

- There is a small amount of variability
  (the IQR is about 1.9kg and Std. Dev. is 1.42kg)

- Is this variability due to subjects fluctuating over time (**within-subject longitudinal effects**) or to (**cross-sectional**) differences between subjects (after adjusting for age)?

- **"Working" Model** for separating within-subject from between-subject variance:

$$Y_{ij} = b_i + \epsilon_{ij}$$

  where Y_{ij} is the weight residualized on age,

b_i and ϵ_{ij} are independent random variables with mean 0,

$$\sigma_b^2 = \mathrm{var}(b_i) = \textbf{between-subj. var.}$$

$$\sigma_w^2 = \mathrm{var}(\epsilon_{ij}) = \textbf{within-subj. var.}$$

$$\sigma_t^2 = \mathrm{var}(Y_{ij}) = \sigma_b^2 + \sigma_w^2 = \textbf{total var.}$$

("Components of variance" decomposition)

- **Note:** There is no **time factor** (age) in this model because we are exploring residuals having removed the effects of time (age)

- **Proportion of total variance** due to between-subject fluctuations is

$$\rho = \sigma_b^2/\sigma_t^2$$

- ρ is also the **within-subject correlation**
  (i.e., correlation between two different observations on same subject):

$$\begin{aligned}
\mathrm{corr}(Y_{ij}, Y_{ij'}) &= \mathrm{corr}(b_i + \epsilon_{ij}, b_i + \epsilon_{ij'}) \qquad (j \neq j') \\
&= \frac{\mathrm{E}\{(b_i + \epsilon_{ij})(b_i + \epsilon_{ij'})\}}{\sqrt{\mathrm{E}(b_i + \epsilon_{ij})^2 \mathrm{E}(b_i + \epsilon'_{ij})^2}} \\
&= \frac{\mathrm{E}(b_i^2)}{\sqrt{\mathrm{var}(Y_{ij})\mathrm{var}(Y_{ij})}} \\
&= \sigma_b^2/\sigma_t^2 \\
&= \rho
\end{aligned}$$

  – b_i and ϵ_{ij} are independent random variables with mean 0
  – ρ is also called ICC (intraclass correlation coefficient)

- The estimated total st. dev. σ_t is: 1.42
- The estimated between-subject st. dev. σ_b is: 1.38
- The estimated within-subject st. dev. σ_w is: 0.37
- The within-subject correlation ρ is 0.91, suggesting that most of the residual variance in weight is due to variance across subjects
  – Subjects are very different from each other $\Rightarrow$ Within each subject, observations are very similar

Summary

In exploratory / descriptive analysis of longitudinal data, we have discussed:

- Exploring basic structure of a longitudinal data set in terms of type of response, number, frequency and timing of measurements, balance and completeness (missing data)
- Displaying population average features of the data by time:

  Time-plots plot the response distribution and mean as a function of time

  These plots are analogous to **marginal models** for longitudinal data because they do not include any longitudinal information
- Adding some longitudinal information to population average plots:

  E.g., subject-level time-plots for a few subjects

  Added trajectories are analogous to **subject specific** models

- Decomposing the data into cross-sectional patterns and longitudinal patterns:
  E.g.:
  - within- and between-subject variance components

The "within" (longitudinal) patterns reflect **subject specific** patterns

Supplement: Estimation of "within" and "between" variance components

- Recall the model:

$$Y_{ij} = b_i + \epsilon_{ij}$$

where b_i and ϵ_{ij} are independent random variables with mean 0,

$$\sigma_b^2 = \mathrm{var}(b_i) = \textbf{between-subj. var.}$$

$$\sigma_w^2 = \mathrm{var}(\epsilon_{ij}) = \textbf{within-subj. var.}$$

$$\sigma_t^2 = \mathrm{var}(Y_{ij}) = \sigma_b^2 + \sigma_w^2 = \textbf{total var.}$$

- For the within-subject variance, an **unbiased** estimator is

$$\widehat{\sigma}_w^2 = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2}{N - m} = \frac{SSW}{N - m}$$

where m is the number of subjects, $N = \sum_i n_i$ is the number of observations, and $\bar{y}_i = (\sum_j y_{ij})/n_i$ is the within-subject mean

- Why divide by $(N - m)$ in $\widehat{\sigma}_w^2$?

For a given subject (conditional on b_i)

$$\mathrm{E}\left\{\sum_j (Y_{ij} - \bar{Y}_i)^2\right\} = (n_i - 1)\sigma_w^2,$$

(You should have learned this unbiased variance estimator in other STAT course)

so therefore adding up over subjects,

$$\mathrm{E}\left\{\sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2\right\} = \sum_i (n_i - 1)\sigma_w^2 = (N - m)\sigma_w^2$$

By setting $\sum_i \sum_j (y_{ij} - \bar{y}_i)^2$ to be $\mathrm{E}\{\sum_i \sum_j (Y_{ij} - \bar{Y}_i)^2\}$, and solving, we obtain unbiased estimator

$$\widehat{\sigma}_w^2 = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2}{N - m}$$

- For the total variance, the usual variance estimator (ie, sample variance) for independent data is:

$$\widehat{\sigma}_t^2 = \frac{\sum_i \sum_j (y_{ij} - \bar{y})^2}{N - 1}$$

where N is the total number of observations and $\bar{y}$ is the total mean.

This estimator for σ_t^2 is biased

  - it ignores correlation within subject
  - Y_{ij} are independent only for observations from different subjects

- An improved estimator arises by noting that

$$\mathrm{E}\{(Y_{ij} - Y_{i'k})^2\} = 2\sigma_t^2, \quad \forall i \neq i'$$

- This suggests that we can obtain an **unbiased** $\widehat{\sigma}^2_{t(unbiased)}$ as the average all half-squared between-subject differences,

$$\frac{(y_{ij} - y_{i'k})^2}{2},$$

$$(i = 1, \ldots, m-1,\ i' = i, \ldots, m,\ j = 1, \ldots, n_i,\ k = 1, \ldots, n_{i'})$$

- Since $\widehat{\sigma}_w^2$ is unbiased, the **unbiased** between-subject variance estimator is

$$\widehat{\sigma}_{b(unbiased)}^2 = \widehat{\sigma}_{t(unbiased)}^2 - \widehat{\sigma}_w^2$$