Exploring Data Analysis

- **Exploratory analysis**: detective work
- **Confirmatory analysis**: judicial work
- show as much as of the relevant data as possible rather than only data summaries
- highlight aggregate patterns of potential scientific interest
- identify both cross-sectional and longitudinal patterns as in example 1.1
- make easy identification of unusual people or unusual observations

if YOU can't see it, DON'T believe it!

Appropriate EDA techniques

- Lines plots (spaghetti plot)
- Average and distribution plots (boxplot, quantiles)
- Empirical covariance
- Residual "pairs" plot
- Variogram

Displays of the responses against time

- Scatterplot of the response variable against time
- example of the 48 pigs (weights versus time)

Displays of the responses against a covariate

- CD4 + example: depressive symptoms (CESD score) versus capacity of immune response

ZAP-plot

1. regress y_{ij} on t_{ij} and get the residuals r_{ij}
2. choose one dimensional summary of the residuals, for example $g_i = \text{median}(r_{i1}, \ldots, r_{in_1})$
3. plot r_{ij} versus t_{ij} using points
4. order units by g_i
5. add lines for selected quantiles of g_i

Graphical methods to separate CS information from LS information

- $Y_{ij} = \beta_c x_{i1} + \beta_L(x_{ij} - x_{i1}) + \epsilon_{ij},\ i = 1, \dots, m,\ j = 1, \dots, n$

  this model implies two facts:

1. $Y_{i1} = \beta_c x_{i1} + \epsilon_{i1},\ i = 1, \dots, m$
2. $Y_{ij} - Y_{i1} = \beta_L(x_{ij} - x_{i1}) + \epsilon_{ij} - \epsilon_{i1}$

   this suggest making two scatterplots

1. y_{i1} against x_{i1} for $i = 1, \dots, m$
2. $y_{ij} - y_{i1}$ against $x_{ij} - x_{i1}$ for $i = 1, \dots, m,\ j = 1, \dots, n$

Fitting smooth curves to longitudinal data

Non parametric regression models that can be used to estimate the mean response profile as a function of time

- Data $(y_i, t_i),\ i = 1, \dots, m$
- we want to estimate an unknown mean response curve $\mu(t)$ in the underlying model

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

- **Kernel estimation**
- **Smoothing Spline**
- **Loess**

Kernel estimation:

- selection of window 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
- weighting function that changes smoothly with time and gives weights to the observations closer to t. Gaussian kernel $K(u) = \exp(0.5u^2)$

- **Smoothing spline:**
- Is the function $s(t)$ which minimizes the criterion $J(\lambda) = \sum_{i=1}^{m} \{y_i - s(t_i)\}^2 + \lambda \int s^{''}(t)^2 dt$
- $s(t)$ satisfy the criterion if and only if is a piece-wise cubic polynomial
- **Loess:**

1. center a window at time t_i
2. fit weighted least squares
3. calculate the residuals (vertical distance from the fitted line to each point in the window)

- down weight the outliers and repeat 1,2,3 many times
- the result is a fitted line that is insensitive to the observations with outlying Y values

Exploring correlation structure

- Regress y_{ij} on x_{ij} to obtain residuals
$$r_{ij} = y_{ij} - \hat{\beta} x_{ij}$$
- with data collected at fixed numbers of equally spaced points, correlation can be studied using scatterplot matrix in which r_{ij} is plotted against r_{ik} for $j < k$.
- Def: if residuals have constant mean and variance and if $corr(y_{ij}, y_{ik})$ depends only on $\mid t_{ij} - t_{ik} \mid$ then the process Y_{ij} is said to be *weakly stationary*
- Scatterplot matrix of CD4+ residuals

Autocorrelation function

- Assuming stationarity, a single correlation estimate can be obtained for each distinct values of the time separation or lag $u = \mid t_{ij} - t_{ik} \mid$. This corresponds to pooling observations pairs along the diagonals of the scatterplot matrix.
- $\rho(u) = Corr(Y_{ij}, Y_{ij-u})$
- standard error of $\rho(u)$ is roughly $1/\sqrt{N}$ where N is the number of independent pairs of observations in the calculation.
- The autocorrelation function is most effective for studying equally spaced data that are roughly stationary.

Autocorrelation function

- Autocorrelation function is most effective for studying equally spaced data that are roughly stationary
- Autocorrelations are more difficult to estimate with irregularly spaced data unless we round observations times as was done for the CD4 data

Variogram

An alternative function describing associations among repeated observations with irregular observation times is the *Variogram* so defined:
$$\gamma(u) = \frac{1}{2} E\left[\{Y(t) - Y(t-u)\}^2\right],\ u \geq 0$$

- If $Y(t)$ is stationary, the Variogram is directly related to the autocorrelation function $\rho(u)$, by
$$\gamma(u) = \sigma^2\{1 - \rho(u)\}$$
where σ^2 is the variance of Y.

Computation of Sample Variogram

- Starting with the residuals r_{ij} and the time t_{ij}, compute all possible

$$\begin{aligned} v_{ijk} &= \tfrac{1}{2}(r_{ij} - r_{ik})^2 \text{ and} \\ u_{ijk} &= t_{ij} - t_{ik} \quad \text{for} \quad j < k \end{aligned}$$

- Now smooth v_{ijk} against u_{ijk} (using lowess)
- Estimate the total variance as

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{ij} (r_{ij} - \bar{r})^2$$

Example on the Protein Content of Milk

- First, compute residuals, allowing for a different mean for each time and diet
- The overall variance of the residuals is $0.2942^2 = 0.087$
- There are 19 time points, so there are 18 lags
- **Note:** the horozontal line is $\hat{\sigma}^2$

A General Serial Covariance Model

Diggle (1988) proposed the following model

$$Y_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta} + \alpha_i + W_i(t_{ij}) + \epsilon_{ij}$$

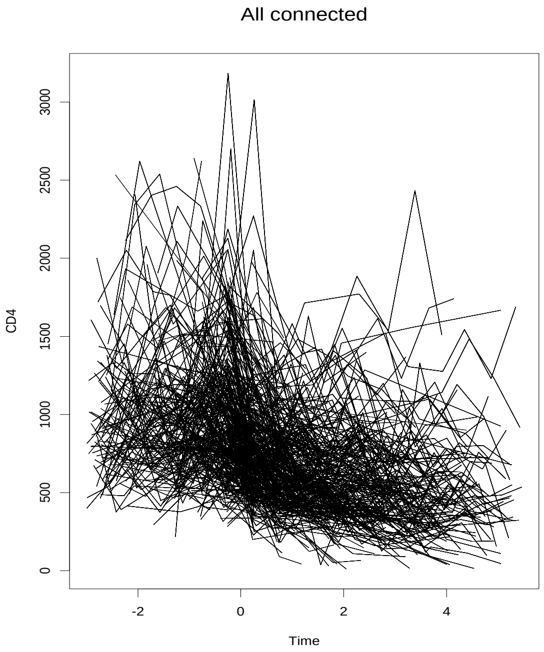
This model contained three sources of variation:

random intercept	α_i
serial process	$W_i(t_{ij})$
measurement error	ϵ_{ij}

If we further assume

$$\begin{aligned} \text{var}(\alpha_i) &= \nu^2 \\ \text{cov}(W(s), W(t)) &= \rho(\mid s - t \mid) \\ \text{var}(\epsilon_{ij}) &= \tau^2 \end{aligned}$$

Then we can use the Variogram to characterize these variance components



5 random individuals

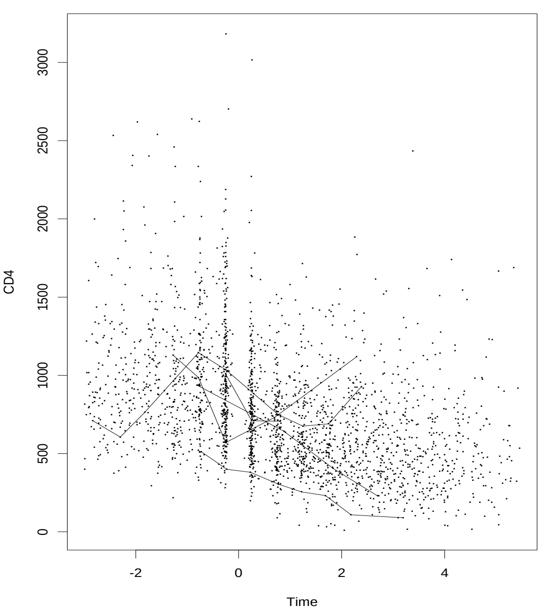
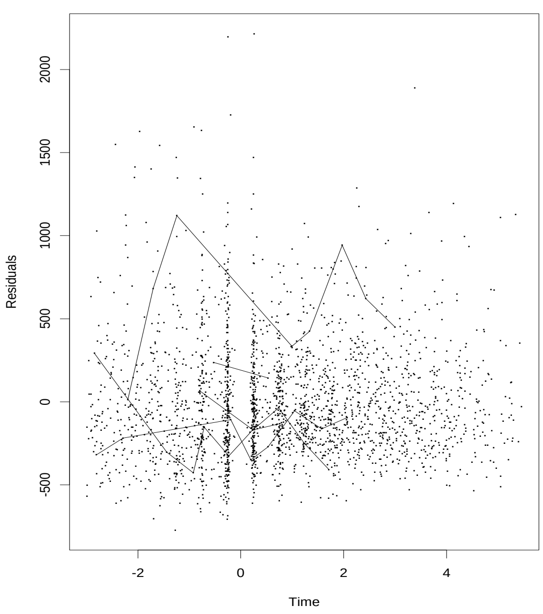


Figure 3.5: Running mean residuals



''Quantiles''

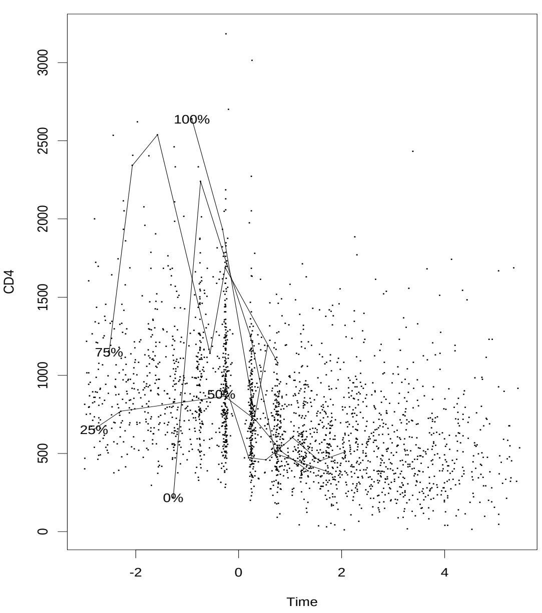


Figure 3.10: Running Mean