General Linear Model for Correlated Data

Objectives:

- Weighted least squares methods
  - Moment estimation
  - Sandwich variance
- Linear mixed models.
  - Models for covariance
  - Maximum likelihood and REML
- Empirical Bayes estimation

General Linear Model for Correlated Data

Example:

- Longitudinal FEV in cystic fibrosis patients.

Example:

- Longitudinal CD4 count in HIV patients.

Example:

- Depression score in a community randomized trial.

Example:

- Lung function, asthma, and air pollution.

General Linear Model for Correlated Data

Consider a sample of N randomly selected units:

$$\boldsymbol{Y}_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \cdots \\ Y_{in_i} \end{pmatrix} \quad i = 1, 2, \ldots, N$$

where the $\boldsymbol{Y}_i$ are independent vectors and n_i may or may not be the same for all units i.

General Linear Model for Correlated Data

Associated with the jth measurement on the ith unit is a $1 \times p$ vector of covariates

$$\underset{(1 \times p)}{\boldsymbol{X}_{ij}} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})$$

$$\underset{(n_i \times p)}{\boldsymbol{X}_i} = \begin{pmatrix} \boldsymbol{X}_{i1} \\ \boldsymbol{X}_{i2} \\ \cdots \\ \boldsymbol{X}_{in_i} \end{pmatrix}$$

In the design matrix $\boldsymbol{X}_i$ the rows correspond to different times of measurement, and the columns are different variables.

Covariates may be:

1. **Cluster specific**, time invariant, or between-subject, so that

$$X_{i1k} = X_{i2k} = \ldots = X_{in_ik}$$

for some $1 \leq k \leq p$. Examples include sex and race in a longitudinal study, and fixed experimental conditions in a longitudinal clinical trial.

2. **Subject specific**, time varying, or within-subject, i.e., covariate k varies with j so that $X_{ijk} \neq X_{ij'k}$. Examples include time since baseline, experimental condition in crossover or repeated measures designs, smoking status or height in a longitudinal study, or individual characteristics in a clustered sample survey. In some cases (pure repeated measures designs, or longitudinal studies with fixed time points), $X_{ijk} = X_{i'jk}$ for all j.

3. **Fixed by design**, e.g., treatment group indicator, time since baseline, or individual characteristics in a sample survey.

4. **Stochastic**, e.g., height, current smoking status, or pollution exposure in a longitudinal survey.

General Linear Model

The model assumes:

1. If $\boldsymbol{X}_i$ is stochastic, $(\boldsymbol{Y}_1, \boldsymbol{X}_1), \ldots, (\boldsymbol{Y}_N, \boldsymbol{X}_N)$ are independently distributed. If $\boldsymbol{X}_i$ is fixed by design, then $\boldsymbol{Y}_i$ are independent.
2. Given $\boldsymbol{X}_i$

$$\begin{array}{rcl} E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) & = & \boldsymbol{X}_i\boldsymbol{\beta} \\ (n_i \times 1) & & (n_i \times p)\ \ (p \times 1) \end{array}$$

$$\begin{array}{rcl} \mathrm{cov}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) & = & \boldsymbol{\Sigma}_i \\ & & (n_i \times n_i) \end{array}$$

General Linear Model

- (1) simply states that the sample consists of independently selected units.

- (2) says that given $\boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots, \boldsymbol{X}_{in_i}$, the mean of Y_{ij} is linear and depends on $\boldsymbol{X}_{ij}$:

$$E(Y_{ij} \mid \boldsymbol{X}_i) = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp}$$

Note:

The model may not hold with certain stochastic time varying covariates. The model implies

$$E(Y_{ij} \mid \boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots, \boldsymbol{X}_{in_i}) = E(Y_{ij} \mid \boldsymbol{X}_{ij})$$

but if current outcomes predict future values of the covariates then the mean of the outcome at a given occasion may depend on future covariates. For example, if Y_{ij} is a symptom measure, and X_{ij} is an indicator of drug treatment then past symptoms may influence current treatment (usually a good idea!).

Formally,

$$\begin{aligned} \text{if} \quad & f(X_{i,j+1} \mid Y_{ij}, X_{i,j}) \quad \neq \quad f(X_{i,j+1} \mid X_{ij}) \\ \text{then} \quad & f(Y_{ij} \mid X_{i,j}, X_{i,j+1}) \quad \neq \quad f(Y_{ij} \mid X_{i,j}) \end{aligned}$$

So that the conditional expectation $E(Y_{ij} \mid \boldsymbol{X}_i)$ may not be correct.

Note:

The covariance $\mathbf{\Sigma}_i$ allows for dependencies among measurements on the same unit. Covariance may vary with covariates, e.g., across treatment group, or the covariance may be a function of time. Alternatively, $\mathbf{\Sigma}_i$ may depend on i only through n_i.

Covariance Matrix

We will consider two approaches where $\mathbf{\Sigma}_i$ is unstructured (only for "balanced" data), and where $\mathbf{\Sigma}_i$ has a specified structure.

Define: A **Balanced** and **complete** design means that all subjects are measured at the same n occasions. Balanced only means that subjects should be measured at the same occasions, but some subjects are not observed at all occasions ($n_i < n$).

GLMCD using stacked notation

The GLMCD can be written as:

$$\begin{array}{rcll} E(\boldsymbol{Y} \mid \boldsymbol{X}) & = & \boldsymbol{X}\boldsymbol{\beta} & \\ (\sum_i n_i \times 1) & & (\sum_i n_i \times p) & (p \times 1) \end{array}$$

$$\text{cov}(\boldsymbol{Y} \mid \boldsymbol{X}_i) = \begin{pmatrix} \boldsymbol{\Sigma}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{\Sigma}_2 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & \boldsymbol{\Sigma}_n \end{pmatrix}$$

- If Balanced and complete then $\sum_i n_i = n \cdot N$.

Examples

⋆ One sample repeated measures ANOVA

N subjects are measured repeatedly under n different experimental conditions. The goal is to quantify differences in experimental conditions.

$$E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_n \end{pmatrix}$$

Here $\boldsymbol{X}_i = \boldsymbol{I}_n$, $\boldsymbol{\beta} = \boldsymbol{\mu}$, and $p = n$.

Example: Repeated Measures ANOVA

It is often assumed that the covariance of $\boldsymbol{Y}_i$ has a **compound symmetric** form which arises from the model:

$$Y_{ij} = \mu_j + \alpha_i + \epsilon_{ij}$$

where the α_i's and the ϵ_{ij}'s are independent of each other, with $\text{var}(\alpha_i) = \tau^2$ and $\text{var}(\epsilon_{ij}) = \sigma^2$.

Example: Repeated Measures ANOVA

We can use vector notation to represent the model as:

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \alpha_i \mathbf{1} + \boldsymbol{\epsilon}_i$$

$$\text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \tau^2 \mathbf{1}\mathbf{1}^T + \sigma^2 \boldsymbol{I}_n$$

$$= \begin{pmatrix} \sigma^2+\tau^2 & \tau^2 & \dots & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \dots & \tau^2 \\ & & \ddots & \\ \tau^2 & \tau^2 & \dots & \sigma^2+\tau^2 \end{pmatrix}$$

Examples

⋆ One way multivariate ANOVA (MANOVA)

We assume G treatment groups, and n measurements are obtained in each of N_g subjects in treatment group g.

Goal is to test if the mean vector is the same for all G groups, where we assume the mean model:

$$\begin{array}{rcll} E(\boldsymbol{Y}_{ig} \mid \boldsymbol{X}_i) & = & \boldsymbol{\mu}_g & g = 1, 2, \ldots, G \\ (n \times 1) & & (n \times 1) & \end{array}$$

Example: MANOVA

$$E(\boldsymbol{Y}_{ig} \mid \boldsymbol{X}_i) \quad = \quad (\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{I}_n, \mathbf{0} \ldots, \mathbf{0}) \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \ldots \\ \boldsymbol{\mu}_G \end{pmatrix}$$

The usual MANOVA assumes $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ is unstructured:

$$\mathsf{cov}(\boldsymbol{Y}_{ig}) \quad = \quad \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2n} \\ & & \ddots & \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_{nn} \end{pmatrix}$$

Examples

⋆ One group polynomial growth curve model

N subjects are observed at the same times $t_1, t_2, \ldots, t_n$; for example a group of children from the same cohort is observed yearly at ages 6, 7, 8, ..., 12. A linear model of the average response can be expressed as a polynomial in t_j

$$E(Y_{ij} \mid \boldsymbol{X}_i) = \beta_0 + \beta_1 t_j + \beta_2 t_j^2$$

$$E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) = \begin{pmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ & \cdots & \\ 1 & t_n & t_n^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

One model for $\boldsymbol{\Sigma}_i$ arises from assuming that each subject has his or her own growth curve with parameters $\boldsymbol{\beta}_i$:

$$\begin{aligned} E(\boldsymbol{Y}_i \mid \boldsymbol{\beta}_i, \boldsymbol{X}_i) &= \boldsymbol{X}\boldsymbol{\beta}_i \\ \text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{\beta}_i, \boldsymbol{X}_i) &= \sigma^2 \boldsymbol{I}_n \end{aligned}$$

and

$$\begin{aligned} E(\boldsymbol{\beta}_i \mid \boldsymbol{X}_i) &= \boldsymbol{\beta} \\ \text{cov}(\boldsymbol{\beta}_i \mid \boldsymbol{X}_i) &= \boldsymbol{D} \end{aligned}$$

then

$$\begin{aligned} E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) &= \boldsymbol{X}\boldsymbol{\beta} \\ \text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) &= \text{cov}[E(\boldsymbol{Y}_i \mid \boldsymbol{\beta}_i)] + E[\text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{\beta}_i)] \\ \boldsymbol{\Sigma}_i &= \boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T + \sigma^2 \boldsymbol{I}_n \end{aligned}$$

Examples

⋆ Family studies: Hypertension

Here i indexes *family*. For N families the outcome is blood pressure and the covariates include age, gender, weight, height, physical activity, diet, smoking, etc. We include all known risk factors in the mean model, and then study the residual correlation.

$\mathbf{\Sigma}_i$ may be structured for genetic models as follows:

$$\begin{array}{c} M \\ F \\ C_1 \\ C_2 \\ C_3 \end{array} \begin{pmatrix} \sigma_A^2 & & & & \\ \sigma_{MF} & \sigma_A^2 & & & \\ \sigma_{CP} & \sigma_{CP} & \sigma_C^2 & & \\ \sigma_{CP} & \sigma_{CP} & \sigma_{CC} & \sigma_C^2 & \\ \sigma_{CP} & \sigma_{CP} & \sigma_{CC} & \sigma_{CC} & \sigma_C^2 \end{pmatrix}$$

Longitudinal Studies

In many cases the primary focus of a study is the change in the mean response over time. This can be modelled by $\boldsymbol{X}_i\boldsymbol{\beta}$ and then $\boldsymbol{\Sigma}_i$ represents parameters of secondary interest (or quite possible nuisance parameters).

If n_i is large and/or the design is inherently unbalanced then it may be desirable to impose some strucure on $\boldsymbol{\Sigma}_i$.

Method 1: Random effects models

Method 2: Serial correlation models

Banded: When measurements are equally spaced one assumption is that the correlation only depends on the distance

$$\mathbf{\Sigma}_i = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{(n-1)} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{(n-2)} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{(n-3)} \\ & & & \ddots & \\ \rho_{(n-1)} & \rho_{(n-2)} & \rho_{(n-3)} & \cdots & 1 \end{pmatrix}$$

This implies

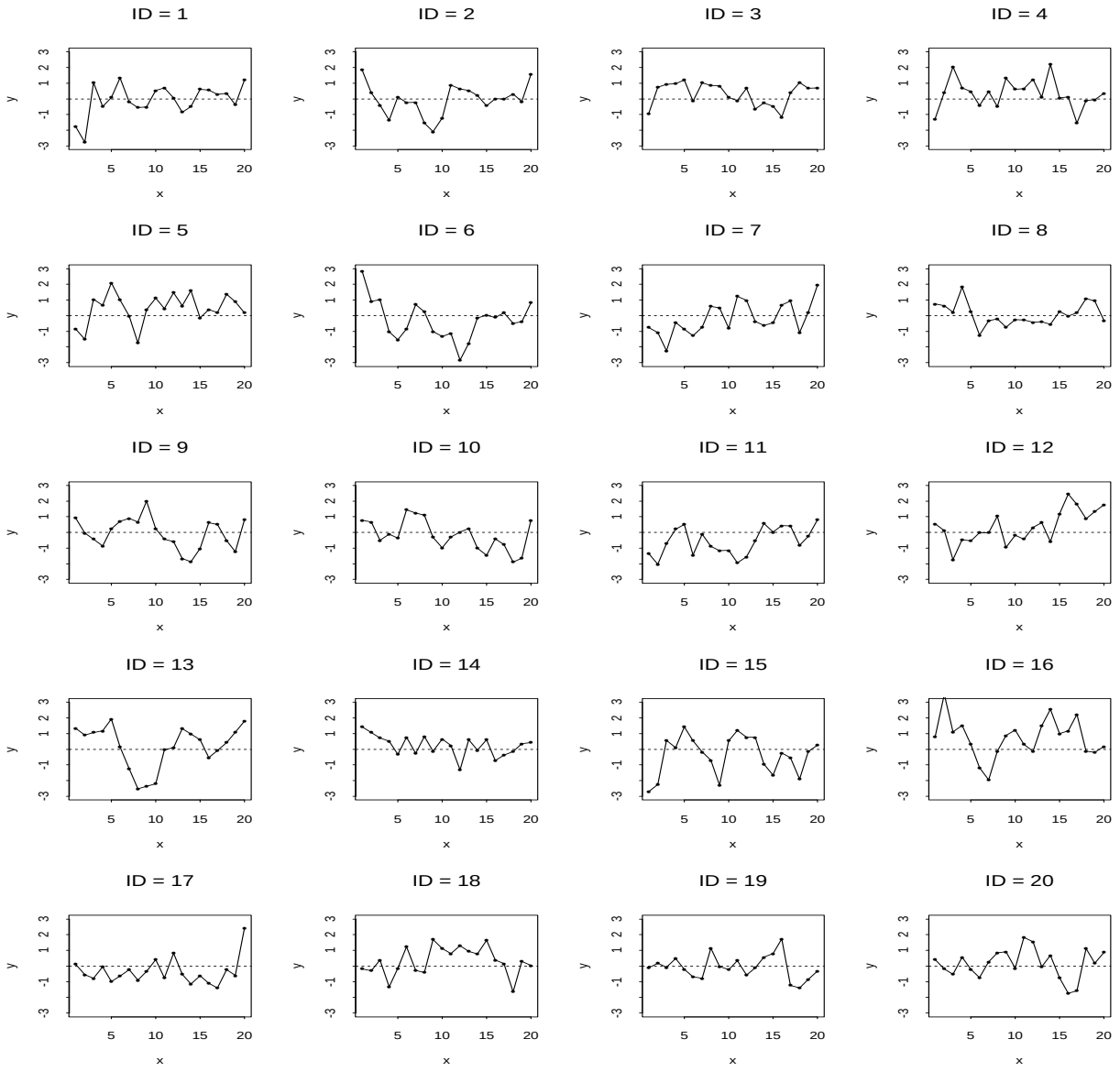
$$\begin{aligned} \text{var}(Y_{ij}) &= \sigma^2 \\ \text{corr}(Y_{i,j}, Y_{i,j+k}) &= \rho_k \quad \forall j, k \end{aligned}$$

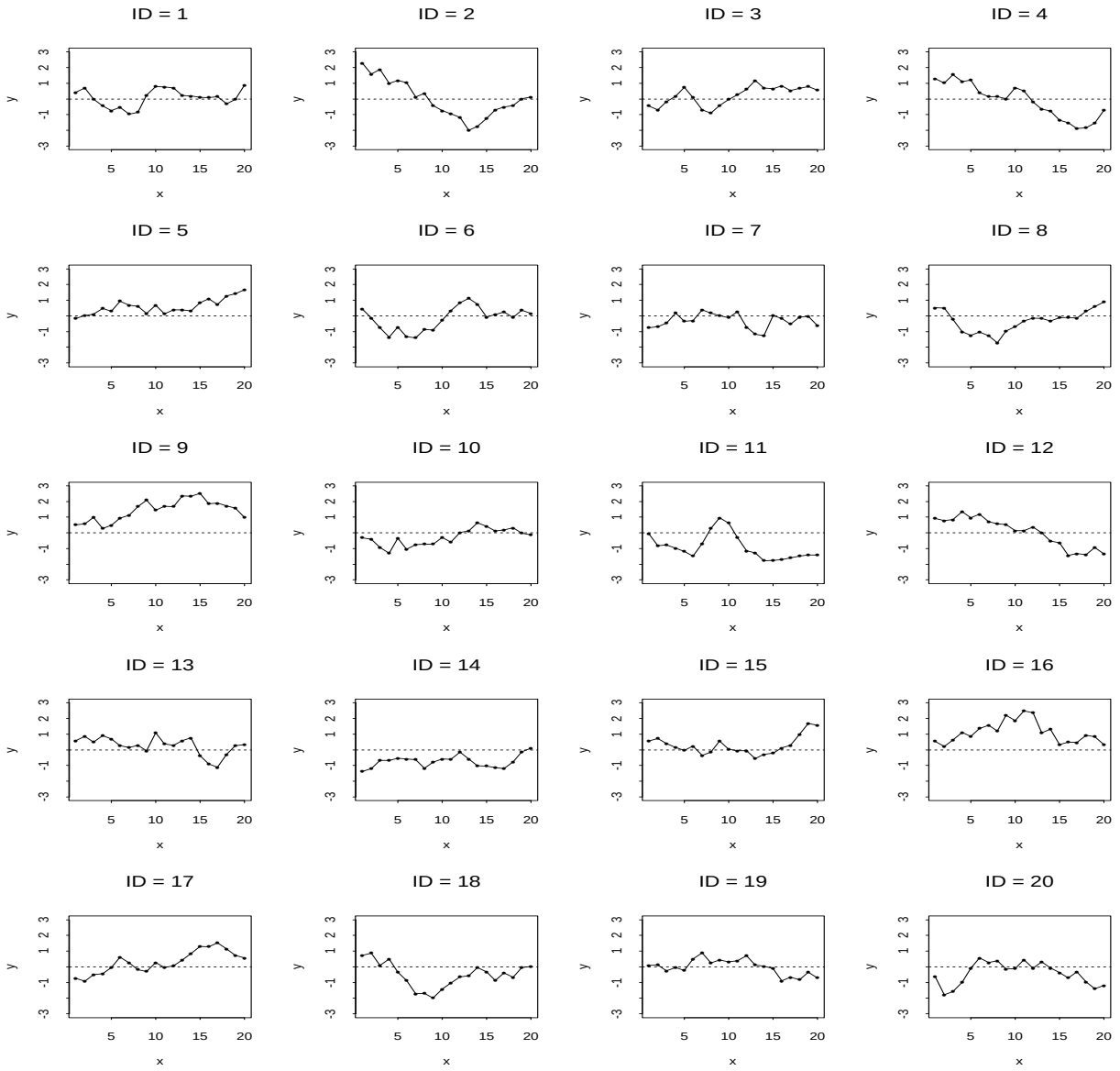
Autoregressive: For data that arise over time it is often reasonable to assume that correlation between measurements that are close in time is greater than the correlation of measurements that are widely separated in time. One model is (here for unit spaced observations):

$$\boldsymbol{\Sigma}_i = \sigma^2 \begin{pmatrix} 1 & \rho^1 & \rho^2 & \dots & \rho^{(n-1)} \\ \rho^1 & 1 & \rho^1 & \dots & \rho^{(n-2)} \\ \rho^2 & \rho^1 & 1 & \dots & \rho^{(n-3)} \\ & & & \ddots & \\ \rho^{(n-1)} & \rho^{(n-2)} & \rho^{(n-3)} & \dots & 1 \end{pmatrix}$$

One construction is given by

$$\begin{aligned} \text{var}(Y_{i0}) &= \sigma^2 \\ Y_{i,j+1} \mid Y_{ij} &= \rho Y_{ij} + \epsilon_{ij} \\ \text{var}(\epsilon_{ij}) &= \sigma^2(1-\rho^2) \quad \text{independent} \end{aligned}$$


rho = 0.4
ID = 1
ID = 2
ID = 3
ID = 4
ID = 5
ID = 6
ID = 7
ID = 8
ID = 9
ID = 10
ID = 11
ID = 12
ID = 13
ID = 14
ID = 15
ID = 16
ID = 17
ID = 18
ID = 19
ID = 20
y
x
-3 -1 1 2 3
5 10 15 20


rho = 0.9
ID = 1
ID = 2
ID = 3
ID = 4
ID = 5
ID = 6
ID = 7
ID = 8
ID = 9
ID = 10
ID = 11
ID = 12
ID = 13
ID = 14
ID = 15
ID = 16
ID = 17
ID = 18
ID = 19
ID = 20
y
x
-3 -1 1 2 3
5 10 15 20

Cross-sectional versus Longitudinal Effects

In our simple growth curve model we assumed that all subjects were measured at the same times, t_j, and were from the same cohort (i.e. same age at baseline). This is rarely the case in observational studies.

Individuals enter at different ages, and measurements may be taken at different times. This design provides the opportunity to obtain information about differences between **cohorts**, as well as differences due to **aging**.

Cross-sectional versus Longitudinal Effects

Ware et al. (1990) discuss a study of pulmonary function where PF was measured every three years for baseline and two follow-up visits on a sample of never-smoking adults. One PF measure is FEV1 (forced expiratory volume in 1 second). They found:

Cross-sectional versus Longitudinal Effects

Possible reasons:

- **Cohort Effects** – younger cohorts are exposed to higher levels of pollution.

- **Attrition** – we may only see older subjects that are healthy.

Cross-sectional versus Longitudinal Effects

We can partition age, X_{ij}, into two components:

- **Cross-sectional comparisons**:

$$E(Y_{i1} \mid \boldsymbol{X}_i) = \beta_0 + \beta_C X_{i1}$$

- **Longitudinal comparisons**:

$$E(Y_{ij} - Y_{i1} \mid \boldsymbol{X}_i) = \beta_L (X_{ij} - X_{i1})$$

Putting these two models together we have:

$$E(Y_{ij} \mid \boldsymbol{X}_i) = \beta_0 + \beta_C X_{i1} + \beta_L (X_{ij} - X_{i1})$$

Missing Data Issues

With longitudinal data we must consider the reasons for missing data since the missing data mechanism (MDM) can impact the validity of estimates and tests.

Examples:

1. **Repeated measures experiment** – HIV patients are given an anti-viral therapy and viral load is measured monthly for 6 months. Some subjects do not comply or drop-out.
2. **Validation study** – Food frequency questionnaires (FFQ) are obtained on all subjects; a validation (more costly but accurate instrument) is obtained for a subset only. We may randomly sample subjects for validation **or** select them based on the FFQ data.

3. **Longitudinal study** – Children are measured for FEV through the schools. Children may move in and out of study schools.

4. **Quality of life study** – Many clinical trials now routinely collect self-reported information on quality of life. Patients may be to ill to give an evaluation.

Missing Data Issues

To formulate different missing data mechanisms we introduce additional notation:

$R_{ij} = 1$ if subject i is observed at time j

$R_{ij} = 0$ if subject i is not observed at time j

MCAR Missing completely at random if

$$f(\boldsymbol{R}_i \mid \boldsymbol{Y}_i, \boldsymbol{X}_i, \Psi) = f(\boldsymbol{R}_i \mid \boldsymbol{X}_i, \Psi)$$

This implies that $E(Y_{ij} \mid R_{ij} = 1, \boldsymbol{X}_i) = E(Y_{ij} \mid \boldsymbol{X}_i)$.

Missing Data Issues

MAR Missing at random if

$$f(\boldsymbol{R}_i \mid \boldsymbol{Y}_i^O, \boldsymbol{Y}_i^M, \boldsymbol{X}_i, \Psi) = f(\boldsymbol{R}_i \mid \boldsymbol{Y}_i^O, \boldsymbol{X}_i, \Psi)$$

Here the probability of missing data only depends on the observed values and not the missing values.

Trouble starts here since this implies
$E(Y_{ij} \mid R_{ij} = 1, \boldsymbol{X}_i) \neq E(Y_{ij} \mid \boldsymbol{X}_i)$ (possibly).

NI Non-ignorable if

$$f(\boldsymbol{R}_i \mid \boldsymbol{Y}_i^O, \boldsymbol{Y}_i^M, \boldsymbol{X}_i, \Psi) \quad \text{depends on} \quad \boldsymbol{Y}_i^M$$

Summary

- GLMCD is flexible.
- Covariate models / issues.
- Covariance models.
- Missing data issues.
  - Estimation – semiparametric.
  - Estimation – parametric (Maximum likelihood).

General Linear Model for Correlated Data

⋆ Estimating β with known Σ

Weighted least squares:

In univariate regression, WLS yields estimates of $\boldsymbol{\beta}$ that minimize the objective function

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{N} w_i (Y_i - \boldsymbol{X}_i \boldsymbol{\beta})^2$$

Analogously, the multivariate version of WLS finds the value of the parameter $\boldsymbol{\beta}(W)$ that minimizes

$$Q_W(\boldsymbol{\beta}) = \sum_{i=1}^{N} (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta})^T \boldsymbol{W}_i (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta})$$

where $\boldsymbol{W}_i$ is an $(n_i \times n_i)$ positive definite symmetric matrix.

It's straight forward to see that

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} Q_W(\boldsymbol{\beta}) = -2 \sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{W}_i (\boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta})$$

GLMCD: WLS

The solution to the minimization solves $\boldsymbol{U}(\boldsymbol{\beta}) = 0$ and yields

$$\widehat{\boldsymbol{\beta}}(W) = \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{Y}_i \right)$$

Example 1:

When $\boldsymbol{W}_i^{-1} = \sigma^2 \boldsymbol{I}_{n_i}$ then

$$Q_I(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{1}{\sigma^2} (Y_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta})^2$$

and $\widehat{\boldsymbol{\beta}}(I)$ is the OLS estimator assuming observations are independent both within and between clusters.

GLMCD: WLS

Example 2:

If $\boldsymbol{X}_i = \boldsymbol{X}$ and $\boldsymbol{W}_i = \boldsymbol{W}$ for all i (e.g. complete and balanced polynomial growth curve data) then,

$$\widehat{\boldsymbol{\beta}}(W) = \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W} \frac{1}{N} \sum_i \boldsymbol{Y}_i$$

This implies that $\widehat{\boldsymbol{\beta}}$ is the regression of the averages.

(**Q**: Is it also the average of the regressions?).

Properties of $\beta(W)$

Given $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_N$ and $\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots \boldsymbol{W}_N$

$$\begin{aligned} E\left[\widehat{\boldsymbol{\beta}}(W)\right] &= \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{W}_i E[\boldsymbol{Y}_i]\right) \\ &= \boldsymbol{\beta} \end{aligned}$$

$$\text{var}\left[\widehat{\boldsymbol{\beta}}(W)\right] = \boldsymbol{A}^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{\Sigma}_i \boldsymbol{W}_i \boldsymbol{X}_i\right) \boldsymbol{A}^{-1}$$

$$\text{where} \quad \boldsymbol{A}^{-1} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{X}_i\right)^{-1}$$

Thus,

$$\boldsymbol{W}_i = \boldsymbol{I}_{n_i} \quad \Rightarrow$$

$$\text{var}\left[\widehat{\boldsymbol{\beta}}(I)\right] = \left(\sum_i \boldsymbol{X}_i^T \boldsymbol{X}_i\right)^{-1} \left(\sum_i \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i \boldsymbol{X}_i\right) \left(\sum_i \boldsymbol{X}_i^T \boldsymbol{X}_i\right)^{-1}$$

$$\boldsymbol{W}_i = \boldsymbol{\Sigma}_i^{-1} \quad \Rightarrow$$

$$\text{var}\left[\widehat{\boldsymbol{\beta}}(\Sigma^{-1})\right] = \left(\sum_i \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i\right)^{-1}$$

Lemma:

$$\text{var}\left[\widehat{\boldsymbol{\beta}}(\boldsymbol{\Sigma}^{-1})\right] \;\leq\; \text{var}\left[\widehat{\boldsymbol{\beta}}(\boldsymbol{W})\right]$$

Notice that $\boldsymbol{W}_i = \boldsymbol{I}_{n_i}$ gives $\widehat{\boldsymbol{\beta}}_{OLS}$ where all observations are treated as independent (weighted equally). It follows that $E(\widehat{\boldsymbol{\beta}}_{OLS}) = \boldsymbol{\beta}$, but $\widehat{\boldsymbol{\beta}}_{OLS}$ may not be very efficient.

For maximum efficiency we must estimate $\boldsymbol{\Sigma}_i$.

Q: How do we compare "efficiencies"?

Answer: We compare the ratio

$$\text{efficiency} \;=\; \frac{\text{var}\left[\widehat{\boldsymbol{\beta}}(\boldsymbol{\Sigma}^{-1})\right]}{\text{var}\left[\widehat{\boldsymbol{\beta}}(W)\right]}$$

Estimating Σ

In general $\boldsymbol{\Sigma}_i$ is unknown. With balanced and complete data we can construct a simple estimator of $\boldsymbol{\Sigma}$ and use this to obtain $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\Sigma}}^{-1})$.

Lemma:

Under regularity conditions on the covariate space $\boldsymbol{X}_i$, if $\widehat{\boldsymbol{W}}_i$ is a consistent estimator of $\boldsymbol{W}_i$ then $\widehat{\boldsymbol{\beta}}(\boldsymbol{W})$ and $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{W}})$ have the same asymptotic distribution.

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}(\boldsymbol{W}) - \boldsymbol{\beta}\right) \quad \rightarrow \quad \mathcal{N}(\,\mathbf{0},\, \boldsymbol{C}_W\,)$$

$$\sqrt{N}\left(\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{W}}) - \boldsymbol{\beta}\right) \quad \rightarrow \quad \mathcal{N}(\,\mathbf{0},\, \boldsymbol{C}_W\,)$$

where

$$\boldsymbol{C}_W = \lim_{N\to\infty} N\boldsymbol{A}_N^{-1}\left(\sum_i \boldsymbol{X}_i^T\boldsymbol{W}_i\boldsymbol{\Sigma}_i\boldsymbol{W}_i\boldsymbol{X}_i\right)\boldsymbol{A}_N^{-1}$$

$$\boldsymbol{A}_N = \sum_i \boldsymbol{X}_i^T\boldsymbol{W}_i\boldsymbol{X}_i$$

Estimating Σ

In the case where $\text{cov}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}$ for all i, we can construct a consistent estimator of the optimal weight matrix $\boldsymbol{W}_i = \boldsymbol{\Sigma}^{-1}$:

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{Y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}(I_n) \right] \left[\boldsymbol{Y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}(I_n) \right]^T \quad .$$

In fact any consistent estimator of $\boldsymbol{\beta}$ will suffice.

Two-step estimator:

Step 1: Obtain $\widehat{\boldsymbol{\beta}}(I_n)$ and $\widehat{\boldsymbol{\Sigma}}$.

Step 2: Obtain $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\Sigma}}^{-1})$.

Estimating Σ

As we shall show, if these two steps are iterated, with balanced and complete data where $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$, then $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\Sigma}}^{-1})$ and $\widehat{\boldsymbol{\Sigma}}$ are also the MLE's assuming multivariate normality.

We can estimate the asymptotic variance of $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\Sigma}}^{-1})$ using

$$\widehat{C}_{\boldsymbol{\Sigma}^{-1}} = N\left(\sum_{i=1}^{N} \boldsymbol{X}_i^T \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}_i\right)$$

Modelling Σ

In n is large (ie. $n = \dim(\boldsymbol{Y}_i)$) then we estimate a large number of parameters in $\boldsymbol{\Sigma}$. In that case we may require large sample sizes, N, before the distribution of $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\Sigma}}^{-1})$ and $\widehat{\boldsymbol{\beta}}(\boldsymbol{\Sigma}^{-1})$ approximately agree.

Q: Can't we adopt some simple structure for $\boldsymbol{\Sigma}$?

Answer: Yes! With small to moderate samples we may use our substantive knowledge about $\boldsymbol{Y}_i$ and exploratory data analysis to guide selection of a covariance model. This permits $\boldsymbol{\Sigma}$ to be modeled in terms of $\boldsymbol{\theta}$, a smaller number of parameters.

Modelling Σ

For example, we may use the compound symmetric covariance model:

$$\begin{aligned} \text{var}(Y_{ij}) &= \theta_1 \\ \text{cov}(Y_{ij}, Y_{ik}) &= \theta_2 \end{aligned}$$

and we may use simple moment estimators to obtain estimates

$$\widehat{\boldsymbol{\theta}}_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n}\sum_{j=1}^{n}\left(Y_{ij} - \boldsymbol{X}_{ij}\widehat{\boldsymbol{\beta}}\right)^2$$

$$\widehat{\boldsymbol{\theta}}_2 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n(n-1)}\sum_{j\neq k}\left(Y_{ij} - \boldsymbol{X}_{ij}\widehat{\boldsymbol{\beta}}\right)\left(Y_{ik} - \boldsymbol{X}_{ik}\widehat{\boldsymbol{\beta}}\right)$$

Q: What if it's not really compound symmetric?

Answer:

$$\widehat{\theta}_1 \quad \rightarrow \quad \theta_1^* = \frac{1}{n}\sum_j \sigma_j^2$$

$$\widehat{\theta}_2 \quad \rightarrow \quad \theta_2^* = \frac{1}{n(n-1)}\sum_{j \neq k} \sigma_{jk}$$

These are simple **moment** estimators and therefore a (general) WLLN implies that these will converge to their limit mean.

Modelling Σ

Again, we obtain asymptotic normality for the WLS estimator that uses $\widehat{\boldsymbol{W}}_i = \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1}$.

Again, there is no difference (asymptotically) between use of $\widehat{\boldsymbol{W}}_i$ and $\boldsymbol{W}_i = \boldsymbol{\Sigma}(\boldsymbol{\theta}^*)^{-1}$.

Again, in general the asymptotic covariance of $\widehat{\boldsymbol{\beta}}(W_{\widehat{\boldsymbol{\theta}}})$ is given in "sandwich" form:

$$\boldsymbol{A}_N =$$

$$\boldsymbol{B}_N =$$

Another Empirical Sandwich!

We can use the independent replication across subjects to estimate the matrix $\boldsymbol{B}_N$.

$$\boldsymbol{B}_N = \sum_{i=1}^{N} \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{*^{-1}} \text{var}(\boldsymbol{Y}_i) \boldsymbol{\Sigma}_i^{*^{-1}} \boldsymbol{X}_i$$

$$\widehat{\boldsymbol{B}}_N =$$

Note:

- The key property of this estmator is consistency – requires large number of subjects (N).

Testing Hypotheses

Finally, consider testing hypotheses of the form:

$$\begin{array}{rcccc} H_0 & : & \boldsymbol{Q}^T\boldsymbol{\beta} & = & \mathbf{0} \\ & & (q \times p)\ (p \times 1) & & (q \times 1) \end{array}$$

Under the null hypothesis we have (asymptotically):

$$\sqrt{N}\boldsymbol{Q}^T\widehat{\boldsymbol{\beta}}(W) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}^T\boldsymbol{C}_W\boldsymbol{Q})$$

So that we can use

$$N\boldsymbol{Q}^T\widehat{\boldsymbol{\beta}}(W)\left(\boldsymbol{Q}^T\boldsymbol{C}_W\boldsymbol{Q}\right)^{-1}\widehat{\boldsymbol{\beta}}(W)^T\boldsymbol{Q} \sim \chi^2(q)$$

as a general Wald statistic.

General Linear Model

Comments:

1. The theory sketched here can be considered **semi-parametric** in the sense that estimation and inference for a parameter $\boldsymbol{\beta}$ can be achieved based solely on specification of the mean.
2. DHLZ (2002) call $\boldsymbol{W}_i^{-1}$ the "working covariance model" since inference using the sandwich variance estimator doesn't require $\boldsymbol{W}_i = \boldsymbol{\Sigma}_i^{-1}$. The matrix $\boldsymbol{W}_i$ is used to improve efficiency.
3. It is possible to allow $\boldsymbol{\Sigma}_i$ to depend on i – we'll see examples.
4. The estimates of $\boldsymbol{\beta}$ and the variance of $\widehat{\boldsymbol{\beta}}$ outlined above are special cases of GEE.
5. DHLZ take a slightly different approach to specifying $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ by assuming $\text{var}(Y_{ij} \mid \boldsymbol{X}_i) = \sigma^2$ and then $\text{cov}(Y_i \mid \boldsymbol{X}_i) = \sigma^2 \boldsymbol{R}(\boldsymbol{\theta})$

Efficiency and WLS

Q: If using $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}$ is optimal for WLS estimation of $\boldsymbol{\beta}$, then how suboptimal is $\widehat{\boldsymbol{\beta}}_{OLS}$?

Recall:

$$\text{var}(\widehat{\boldsymbol{\beta}}_{opt}) = \left(\sum_i \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right)^{-1}$$

$$\text{var}(\widehat{\boldsymbol{\beta}}_{OLS}) = \boldsymbol{A}^{-1} \left(\sum_i \boldsymbol{X}_i^T \boldsymbol{\Sigma}_i \boldsymbol{X}_i \right) \boldsymbol{A}^{-1}$$

$$\text{where} \quad \boldsymbol{A}^{-1} = \left(\sum_i \boldsymbol{X}_i^T \boldsymbol{X}_i \right)^{-1}$$

See: Bloomfield and Watson (1975); Watson (1967)

Comment on notation here:

$$\boldsymbol{Y} \quad = \quad \text{stack}(\boldsymbol{Y}_i) \quad = \quad \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_n \end{pmatrix}$$

$$\boldsymbol{X} \quad = \quad \text{stack}(\boldsymbol{X}_i) \quad = \quad \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_n \end{pmatrix}$$

Comment on notation here:

$$\boldsymbol{\Sigma} \quad = \quad \text{block}(\boldsymbol{\Sigma}_i) \quad = \quad \begin{pmatrix} \boldsymbol{\Sigma}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{\Sigma}_2 & \dots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \dots & \boldsymbol{\Sigma}_N \end{pmatrix}$$

Efficiency of OLS estimators

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

Theorem:

Let $\boldsymbol{C}\boldsymbol{\beta}$ be an estimable function for the linear model

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

where $E(\boldsymbol{\epsilon}) = 0$, and $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \boldsymbol{\Sigma}$. Then

$$\boldsymbol{\Sigma}\mathcal{M}(\boldsymbol{X}) \;=\; \mathcal{M}(\boldsymbol{X})$$

implies that the BLUE and LS estimators of $\boldsymbol{C}\boldsymbol{\beta}$ are equivalent.

Note: $\mathcal{M}(\boldsymbol{X})$ denotes the column space of $\boldsymbol{X}$.

Proof: Watson (1967)

Efficiency of OLS

Example 1:

$$
\begin{aligned}
N &= 10 \\
n &= 5 \\
t_j &= (-2, -1, 0, 1, 2)
\end{aligned}
$$

$$
E(Y_{ij}) = \beta_0 + \beta_1 t_j
$$

$$
\boldsymbol{\Sigma} = \sigma^2 \{(1-\rho)\boldsymbol{I} + \rho \boldsymbol{J}\}
$$

Then,

$$\begin{aligned}
\boldsymbol{\Sigma x} &= \sigma^2 \left\{ (1-\rho)\boldsymbol{I} + \rho \boldsymbol{J} \right\} \boldsymbol{x} \\
&= \sigma^2 (1-\rho) \cdot \boldsymbol{x} + \rho n \overline{x} \cdot \mathbf{1} \\
&= a \cdot \boldsymbol{x} + b \cdot \mathbf{1} \\
\\
&\in \mathcal{M}(\boldsymbol{X})
\end{aligned}$$

Therefore, $\widehat{\boldsymbol{\beta}}_{OLS} = \widehat{\boldsymbol{\beta}}(\boldsymbol{\Sigma})$.

Efficiency of OLS

Theorem: Rao (1965)

$$\left(\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$$

if and only if

$$\boldsymbol{\Sigma} = \boldsymbol{X}\boldsymbol{\Gamma}\boldsymbol{X}^T + \boldsymbol{Z}\boldsymbol{\Theta}\boldsymbol{Z}^T + \sigma^2\boldsymbol{I}$$

where $\boldsymbol{Z}^T\boldsymbol{X} = \boldsymbol{0}$, and $\boldsymbol{\Gamma}, \boldsymbol{\Theta}, \sigma^2$ are arbitrary.

This result implies that for any balanced random effects model (with conditional independence) we will have $\widehat{\boldsymbol{\beta}}_{OLS} = \widehat{\boldsymbol{\beta}}(\boldsymbol{\Sigma})$.

Efficiency of OLS

Example 2: consider the same mean model as in Example 1 but now assume AR(1) errors:

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ & 1 & \rho & \rho^2 & \rho^3 \\ & & 1 & \rho & \rho^2 \\ & & & 1 & \rho \\ & & & & 1 \end{pmatrix}$$

some algebra yields

$$\begin{aligned}
\text{var}(\widehat{\boldsymbol{\beta}}_{OLS}) &= \sigma^2 \begin{bmatrix} V_{11} & 0 \\ 0 & V_{22} \end{bmatrix} \\
V_{11} &= 0.004(5 + 8\rho + 6\rho^2 + 4\rho^3 + 2\rho^4) \\
V_{22} &= 0.002(5 + 4\rho - \rho^2 - 4\rho^3 - 4\rho^4)
\end{aligned}$$

$$\text{var}[\widehat{\boldsymbol{\beta}}(\boldsymbol{\Sigma})] = \sigma^2 \begin{bmatrix} 0.01(5 - 8\rho + 3\rho^2)^{-1} & 0 \\ 0 & 0.005(5 - 4\rho + \rho^2)^{-1} \end{bmatrix}$$

...and for certain values of ρ we obtain:

ρ	0.1	0.2	0.3	0.4	0.5
$e(\beta_0)$	0.998	0.992	0.983	0.973	0.963
$e(\beta_1)$	0.997	0.989	0.980	0.970	0.962

ρ	0.6	0.7	0.8	0.9	0.99
$e(\beta_0)$	0.955	0.952	0.956	0.970	0.996
$e(\beta_1)$	0.952	0.955	0.952	0.955	0.961

Comparisons of this kind (and earlier results) suggest that in many circumstances the OLS estimator is satisfactory. This is not always the case. Consider...

Efficiency of OLS

Example 3: again, assume that we have AR(1) errors and now assume $n = 3$ and that subjects crossover from treatment A to treatment B (and from B to A). Assume that subjects have observed treatment paths in equals numbers:

$$\begin{array}{ll} AAA & BAA \\ AAB & BAB \\ ABA & BBA \\ ABB & BBB \end{array}$$

This is a form of "**crossover**" design.

Efficiency of OLS: Example 3

Now assume that the predictor of interest is treatment group,

$$x_{ij} = \mathbf{1}(\text{TX}_i = B, \text{at time } j)$$

$$E(Y_{ij} \mid \boldsymbol{X}_i) = \beta_0 + \beta_1 \cdot x_{ij}$$

...now for certain values of ρ we obtain:

ρ	0.1	0.2	0.3	0.4	0.5
$e(\beta_0)$	0.993	0.974	0.946	0.914	0.880
$e(\beta_1)$	0.987	0.947	0.883	0.797	0.692

ρ	0.6	0.7	0.8	0.9	0.99
$e(\beta_0)$	0.846	0.815	0.788	0.766	0.751
$e(\beta_1)$	0.571	0.438	0.297	0.150	0.015

Discuss: Why the efficiency difference now?

EDA and Covariance Models

Q: What are the appropriate EDA techniques for longitudinal data?

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

EDA and Covariance Models

Q: What are some parametric covariance models?

- Mixed model ($\sigma^2 \boldsymbol{I} + \boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^T$)
- Nested models ($b_i + b_{ij} + e_{ijk}$)
- Autoregressive models
- Moving average models
- General isotropic correlation models
- Combined Random effects and Serial (Diggle 1988)

Example: (DLZ Example 1.1) **CD4+ Cell Counts**

- HIV attacks CD4+ cells.

- Data from the MACS study.

- $N = 369$ infected men – incident cases.

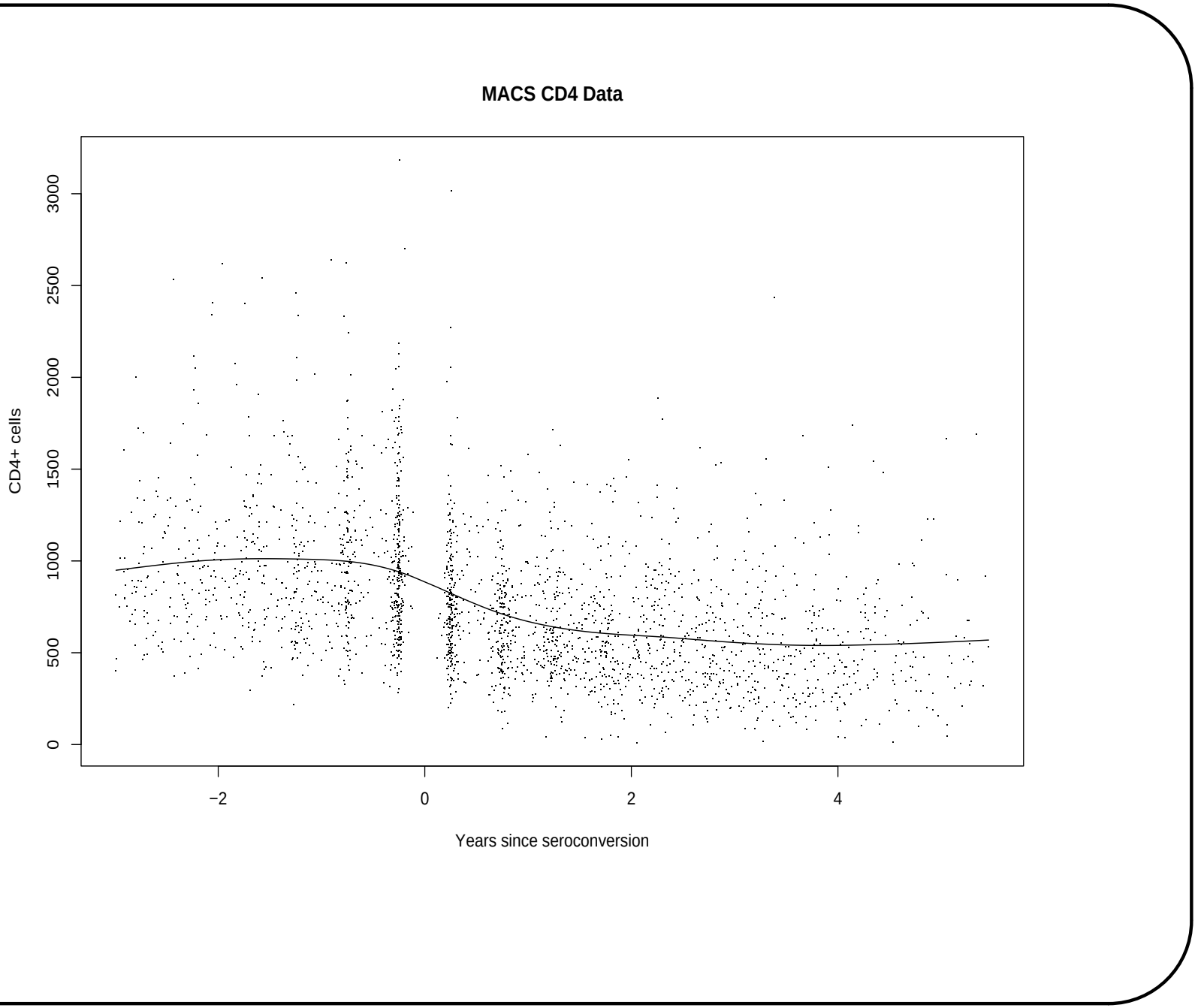
- Analysis focuses on characterizing the process. CD4+ cell counts are used to monitor patient status, and characteristics of the longitudinal process within a patient are thought to be predictive of clinical course. More recently HIV research has focused on longitudinal measures of viral load.

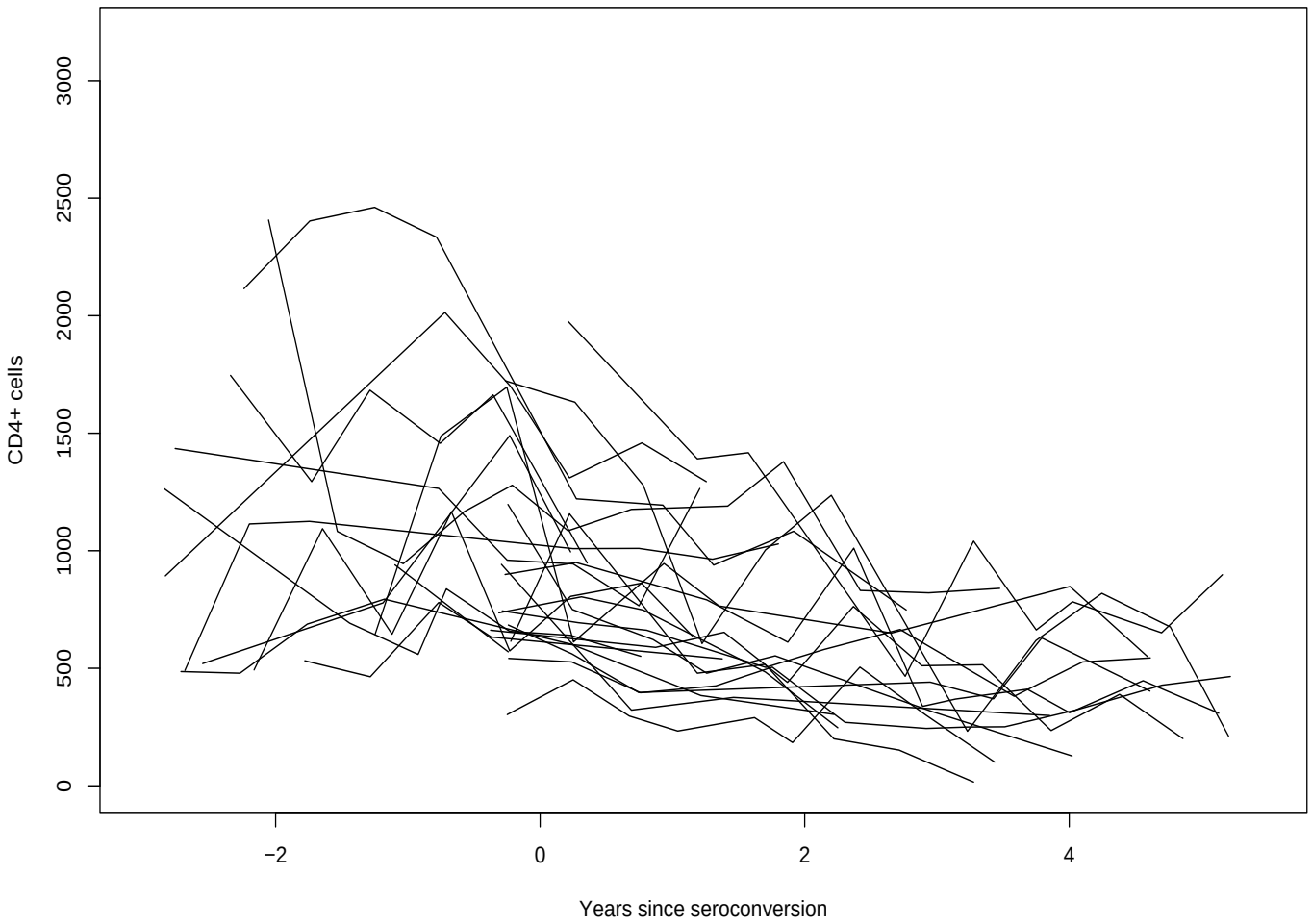
Descriptives:

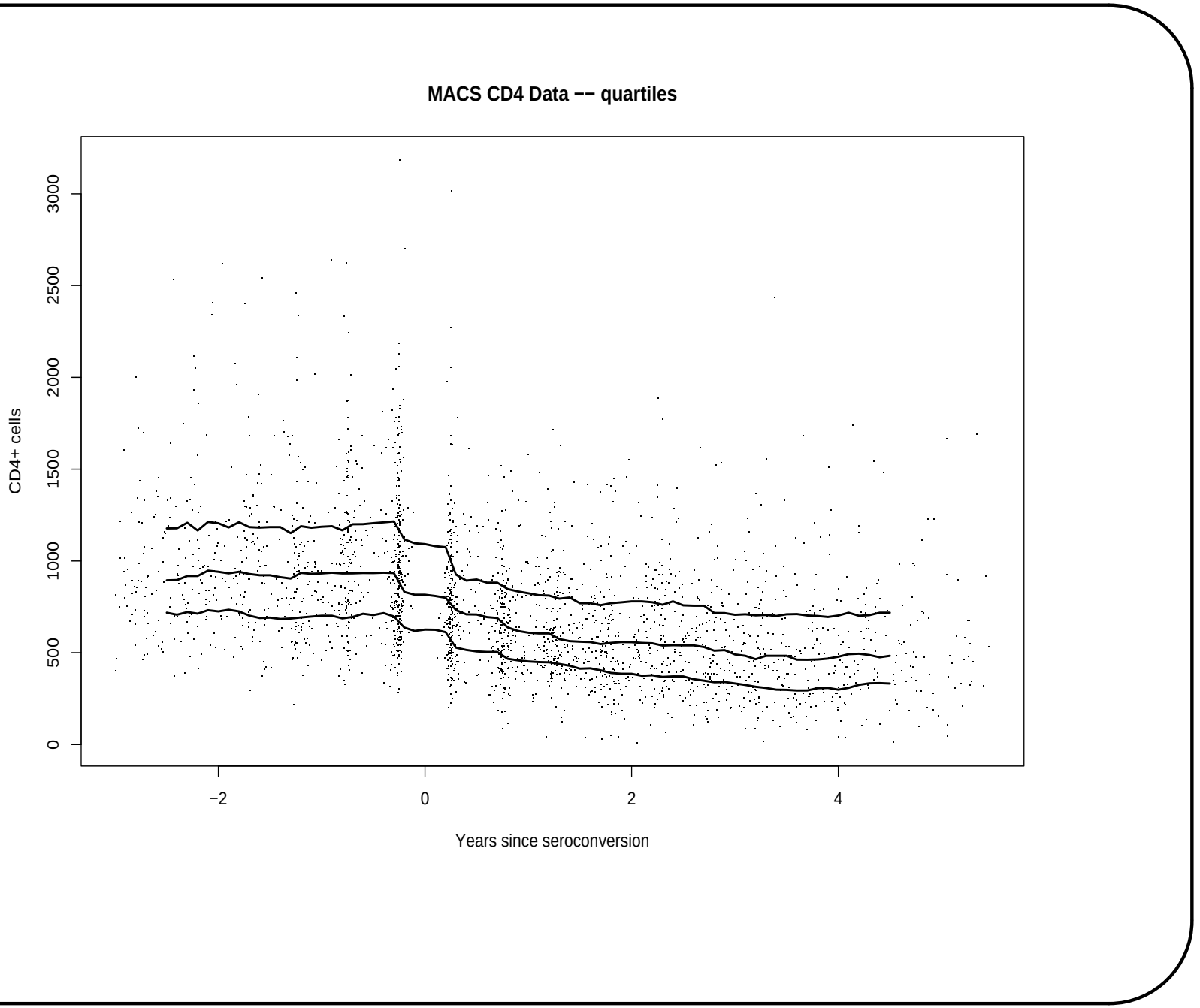
```
      time              cd4              age              packs
 Min.   :-2.9900   Min.   :  10.0   Min.   :-11.290   Min.   :0.0000
 1st Qu.:-0.3922   1st Qu.: 482.8   1st Qu.: -2.760   1st Qu.:0.0000
 Median : 0.7296   Median : 701.5   Median :  1.510   Median :0.0000
 Mean   : 0.8284   Mean   : 765.1   Mean   :  2.636   Mean   :0.9891
 3rd Qu.: 2.1920   3rd Qu.: 964.0   3rd Qu.:  6.950   3rd Qu.:2.0000
 Max.   : 5.4590   Max.   :3184.0   Max.   : 29.080   Max.   :4.0000
     drugs           partners            cesd              id
 Min.   :0.0000   Min.   :-5.00000   Min.   : -7.000   Min.   :10000
 1st Qu.:1.0000   1st Qu.:-3.00000   1st Qu.: -5.000   1st Qu.:11200
 Median :1.0000   Median :-1.00000   Median :  0.000   Median :30050
 Mean   :0.7559   Mean   :-0.03409   Mean   :  2.496   Mean   :26190
 3rd Qu.:1.0000   3rd Qu.: 5.00000   3rd Qu.:  6.000   3rd Qu.:40360
 Max.   :1.0000   Max.   : 5.00000   Max.   : 49.000   Max.   :41840

Number of subjects = 369
Number of observations = 2376

Number of subjects with a given observations-per-subject:
 1  2  3  4  5  6  7  8  9 10 11 12
 5 24 25 47 43 52 40 41 38 21 23 10
```


MACS CD4 Data
CD4+ cells
0
500
1000
1500
2000
2500
3000
−2
0
2
4
Years since seroconversion


MACS CD4 Data −− 25 subjects
CD4+ cells
0
500
1000
1500
2000
2500
3000
−2
0
2
4
Years since seroconversion


MACS CD4 Data −− quartiles
CD4+ cells
0
500
1000
1500
2000
2500
3000
−2
0
2
4
Years since seroconversion

Exploring the Covariance Structure

★ Remove covariate effects first.

```
#####
##### covariance summaries
#####
#
fit <- lm( cd4 ~ ns( year, knots=c(-2,0,2,4) ), data=cd4 )
resids <- cd4$cd4 - fitted( fit )
#
nobs <- length( cd4$cd4 )
nsubjects <- length( table( cd4$id ) )
rmat <- matrix( NA, nsubjects, 7 )
ycat <- c( -2, -1, 0, 1, 2, 3, 4, )
nj <- unlist( lapply( split( cd4$id, cd4$id ), length ) )
for( j in 1:7 ){
     legal <- ( cd4$year >= ycat[j]-0.5 )&( cd4$year < ycat[j]+0.5 )
     jtime <- cd4$year + 0.01*rnorm(nobs)
     t0 <- unlist( lapply(
            split( abs(jtime - ycat[j]) , cd4$id ), min ) )
     tj <- rep( t0, nj )
     keep <- ( abs( jtime - ycat[j] )==tj )&( legal )
```

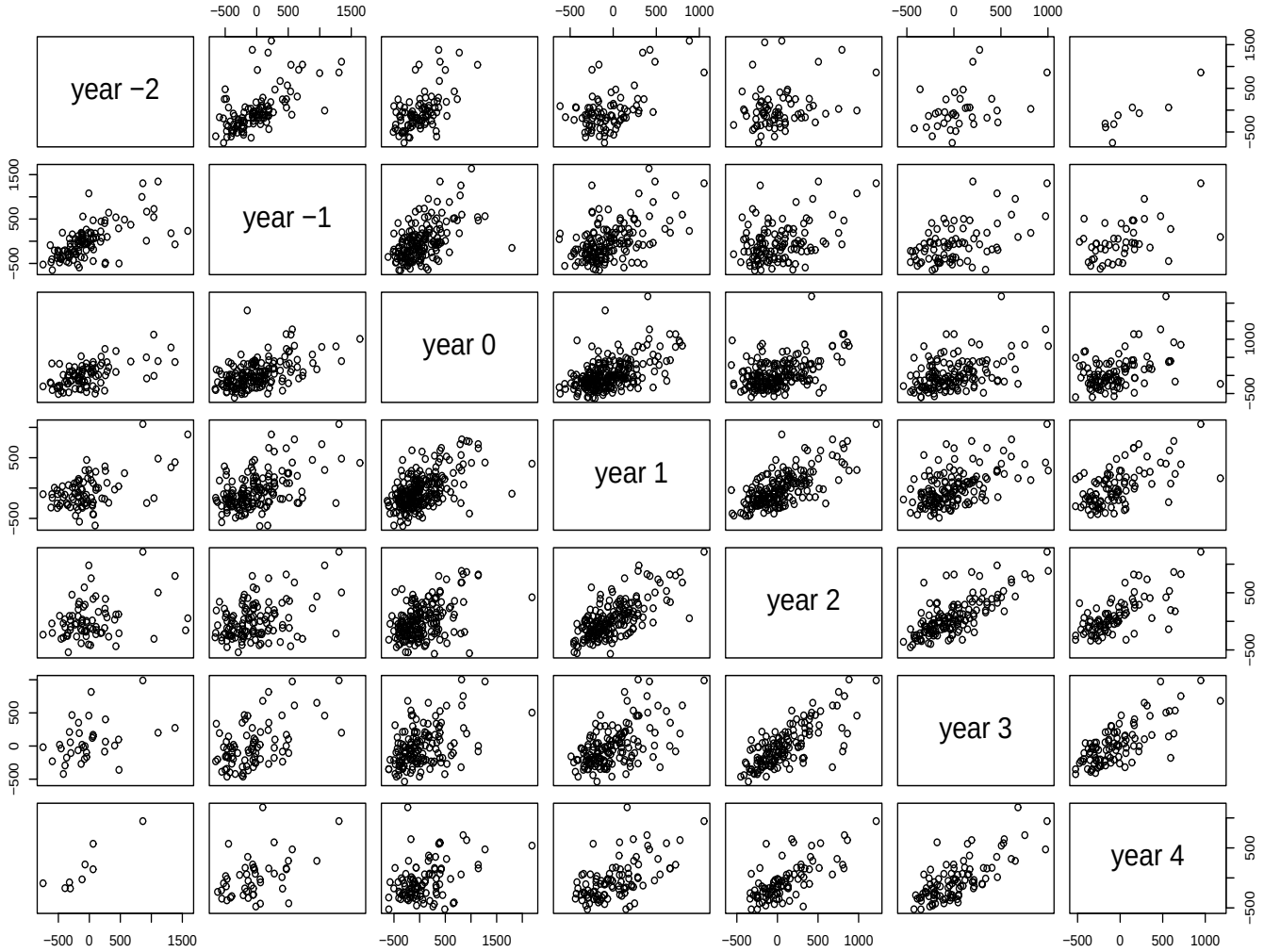
```
      yj <- rep( NA, nobs )
      yj[keep] <- resids[keep]
      yj <- unlist( lapply( split( yj, cd4$id ), min, na.rm=T ) )
      rmat[ , j ] <- yj
}
#
##### covariance matrix
#
cmat <- matrix( 0, 7, 7 )
nmat <- matrix( 0, 7, 7 )
#
for( j in 1:7 ){
  for( k in j:7 ){
        njk <- sum( !is.na( rmat[,j]*rmat[,k] ) )
        sjk <- sum( rmat[,j]*rmat[,k], na.rm=T )/njk
        cmat[j,k] <- sjk
        nmat[j,k] <- njk
  }
}
print( round( cmat, 2 ) )
vvec <- diag(cmat)
cormat <- cmat/( outer( sqrt(vvec), sqrt(vvec) ) )
print( round( cormat, 2 ) )
print( nmat )
```

```
Covariance Matrix
          [,1]      [,2]       [,3]      [,4]      [,5]      [,6]       [,7]
[1,] 195082.2 110392.5  77600.36 66198.26 40901.60 51093.11 117827.90
[2,]      0.0 167084.1  72191.32 61440.78 51217.72 71398.74  63482.42
[3,]      0.0      0.0 140413.48 42555.74 48802.46 41045.85  59231.54
[4,]      0.0      0.0      0.00 80023.49 57418.04 48018.62  54503.60
[5,]      0.0      0.0      0.00     0.00 96497.45 72253.31  63654.61
[6,]      0.0      0.0      0.00     0.00     0.00 90569.13  72262.10
[7,]      0.0      0.0      0.00     0.00     0.00     0.00 101639.49


Correlation Matrix
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]    1 0.61 0.47 0.53 0.30 0.38 0.84
[2,]    0 1.00 0.47 0.53 0.40 0.58 0.49
[3,]    0 0.00 1.00 0.40 0.42 0.36 0.50
[4,]    0 0.00 0.00 1.00 0.65 0.56 0.60
[5,]    0 0.00 0.00 0.00 1.00 0.77 0.64
[6,]    0 0.00 0.00 0.00 0.00 1.00 0.75
[7,]    0 0.00 0.00 0.00 0.00 0.00 1.00
```

Number of observations

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]
[1,]	133	106	91	85	73	35	9
[2,]	0	211	171	157	121	80	46
[3,]	0	0	307	236	192	144	99
[4,]	0	0	0	279	195	149	98
[5,]	0	0	0	0	226	142	89
[6,]	0	0	0	0	0	167	97
[7,]	0	0	0	0	0	0	109


year −2
year −1
year 0
year 1
year 2
year 3
year 4
−500 0 500 1500
−500 0 500 1000

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 contains three sources of random variation:

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

If we further assume

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

Then we can use the **variogram** for EDA.

Variogram

○ For models that have a constant variance the **variogram** is a useful plot. The variogram is defined as:

$$\gamma(u) = \frac{1}{2} E\left[\{Y(t+u) - Y(t)\}^2\right]$$

This function directly relates to the **autocorrelation** function:

$$\begin{aligned} \rho^*(u) &= \text{corr}[Y(t), Y(t+u)] \\ \sigma^2_{\text{Total}} &= \text{var}[Y(t)] \end{aligned}$$

$$\gamma(u) = \sigma^2_{\text{Total}} \{1 - \rho^*(u)\}$$

Note: For the Diggle (1988) model we obtain

$$\gamma(u) = \sigma^2 \{1 - \rho(u)\} + \tau^2$$

EDA for Covariance Structure

Numerical Summaries

- Empirical covariance & correlation

Variogram

Define:

$$\begin{aligned} R_{ij} &= Y_{ij} - \boldsymbol{X}_{ij}\boldsymbol{\beta} \\ &= b_{i,0} + W(t_{ij}) + \epsilon_{ij} \end{aligned}$$

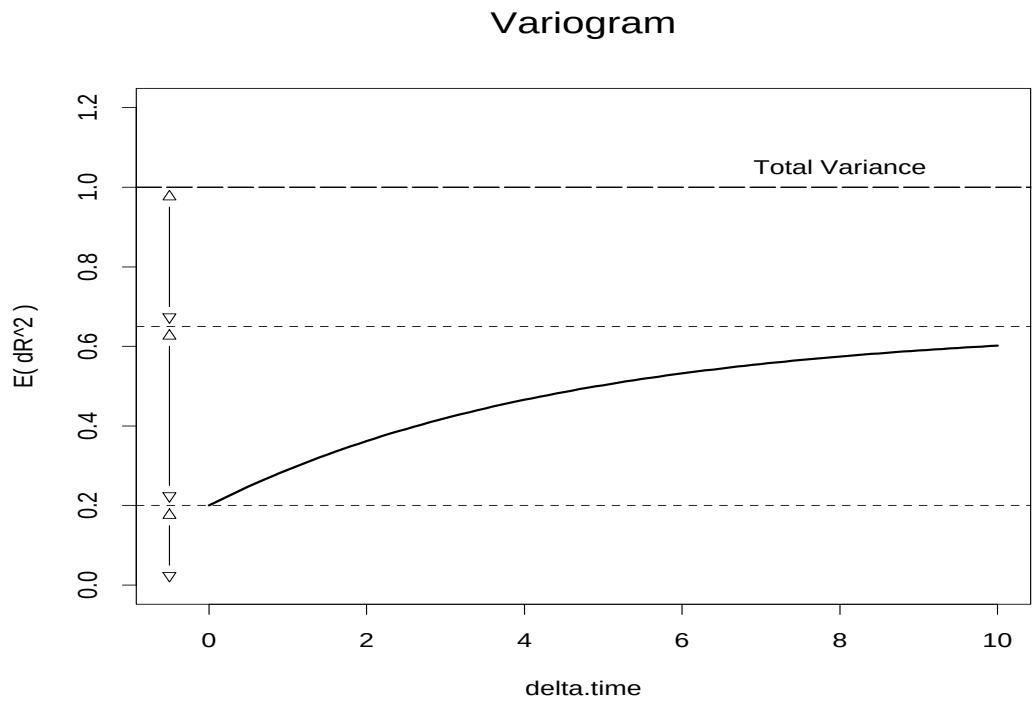
Note:

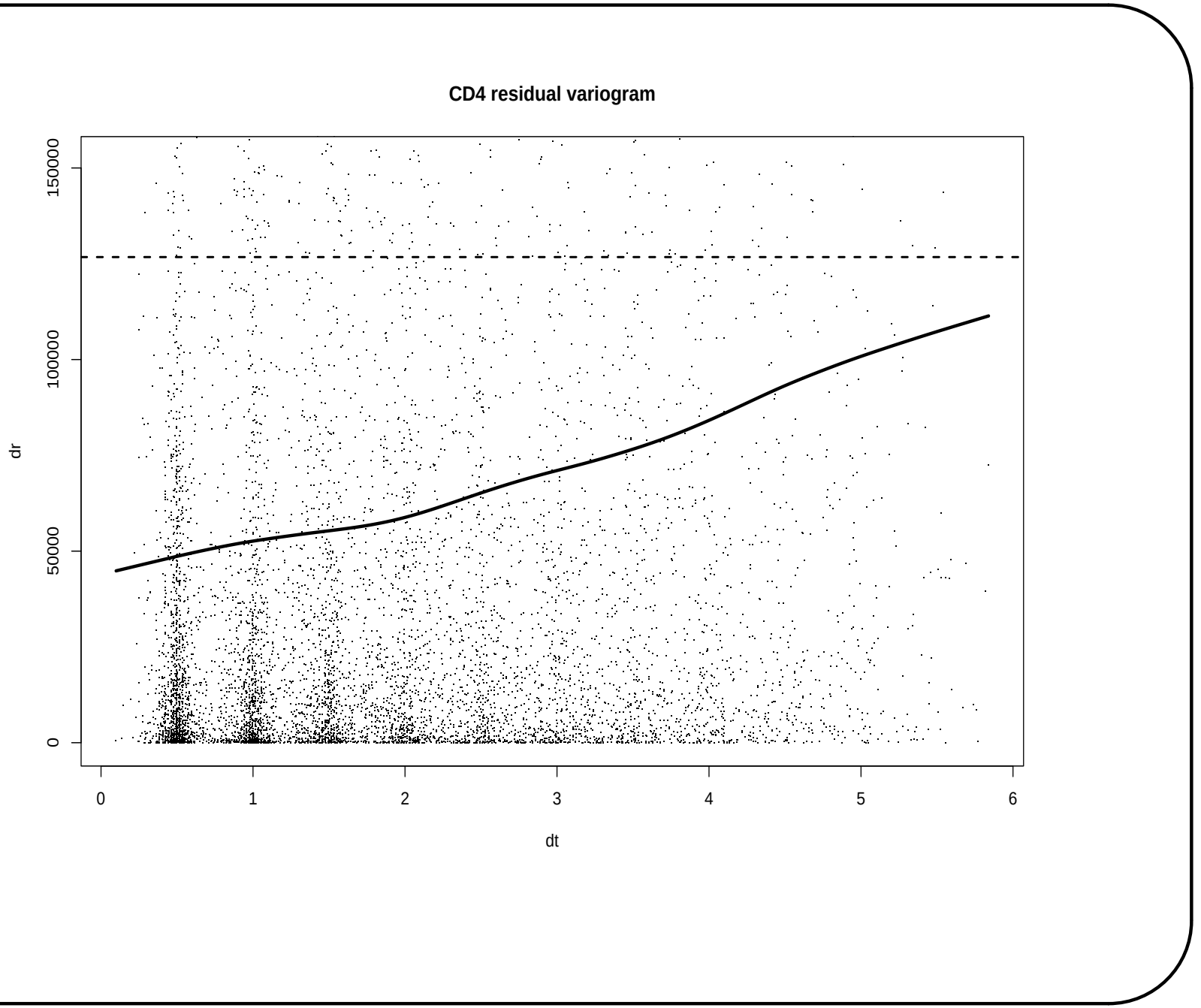
$$\text{var}(R_{ij}) \quad = \quad \nu^2 + \sigma^2 + \tau^2$$

$$E\left[\frac{1}{2}(R_{ij} - R_{ik})^2\right] \quad = \quad \sigma^2 \cdot (1 - \rho^{|t_{ij} - t_{ik}|}) + \tau^2$$

Plot:

$$\frac{1}{2}(R_{ij} - R_{ik})^2 \quad \text{versus} \quad |t_{ij} - t_{ik}|$$


Variogram
Total Variance
E(dR^2)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
2
4
6
8
10
delta.time


CD4 residual variogram
dr
0
50000
100000
150000
0
1
2
3
4
5
6
dt

Variogram

```
#
##### Variogram estimation
#
source("variogram.q")
#
out <- lda.variogram( id=cd4$id, y=resids, x=cd4$year )
dr <- out$delta.y
dt <- out$delta.x
#
var.est <- var( resids )
#
postscript( file="cd4_eda_variogram.ps", horiz=T )
plot( dt, dr, pch=".", ylim=c(0, 1.2*var.est) )
lines( smooth.spline( dt, dr, df=5 ), lwd=3 )
abline( h=var.est, lty=2, lwd=2 )
title("CD4 residual variogram")
graphics.off()
#
```

variogram.q

```
lda.variogram <- function( id, y, x ){
#
# INPUT:  id = (nobs x 1) id vector
#          y = (nobs x 1) response (residual) vector
#          x = (nobs x 1) covariate (time) vector
#
# RETURN:  delta.y = vec( 0.5*(y_ij - y_ik)^2 )
#          delta.x = vec( abs( x_ij - x_ik ) )
#
uid <- unique( id )
m <- length( uid )
delta.y <- NULL
delta.x <- NULL
did <- NULL
for( i in 1:m ){
  yi <- y[ id==uid[i] ]
  xi <- x[ id==uid[i] ]
  n <- length(yi)
  expand.j <- rep( c(1:n), n )
  expand.k <- rep( c(1:n), rep(n,n) )
  keep <- expand.j > expand.k
  if( sum(keep)>0 ){
        expand.j <- expand.j[keep]
        expand.k <- expand.k[keep]
        delta.yi <- 0.5*( yi[expand.j] - yi[expand.k] )^2
```

```
        delta.xi <- abs( xi[expand.j] - xi[expand.k] )
        didi <- rep( uid[i], length(delta.yi) )
        delta.y <- c( delta.y, delta.yi )
        delta.x <- c( delta.x, delta.xi )
        did <- c( did, didi )
  }
}
out <- list( id = did, delta.y = delta.y, delta.x = delta.x )
out
}
#-------------------------------------------------
```

Summary

Basic Data Summaries

- Number of subjects; Number of observations / subject
- Univariate summaries for each variable

EDA for Systematic Variation

- Mean response by covariates
- (Covariate between- and within- variation)

EDA for Random Variation

- Individual plots
- Empirical covariance & correlation
- Variogram if constant variance