Analysis of Longitudinal Data

- Patrick J. Heagerty PhD
- Department of Biostatistics
- University of Washington

Session Three Outline

- Role of correlation
  - ▷ Impact proper standard errors
  - ▷ Used to weight individuals (clusters)

- Models for correlation / covariance
  - ▷ Regression: Group-to-Group variation
  - ▷ Random effects: Individual-to-Individual variation
  - ▷ Serial correlation: Observation-to-Observation variation

Longitudinal Data Analysis

INTRODUCTION

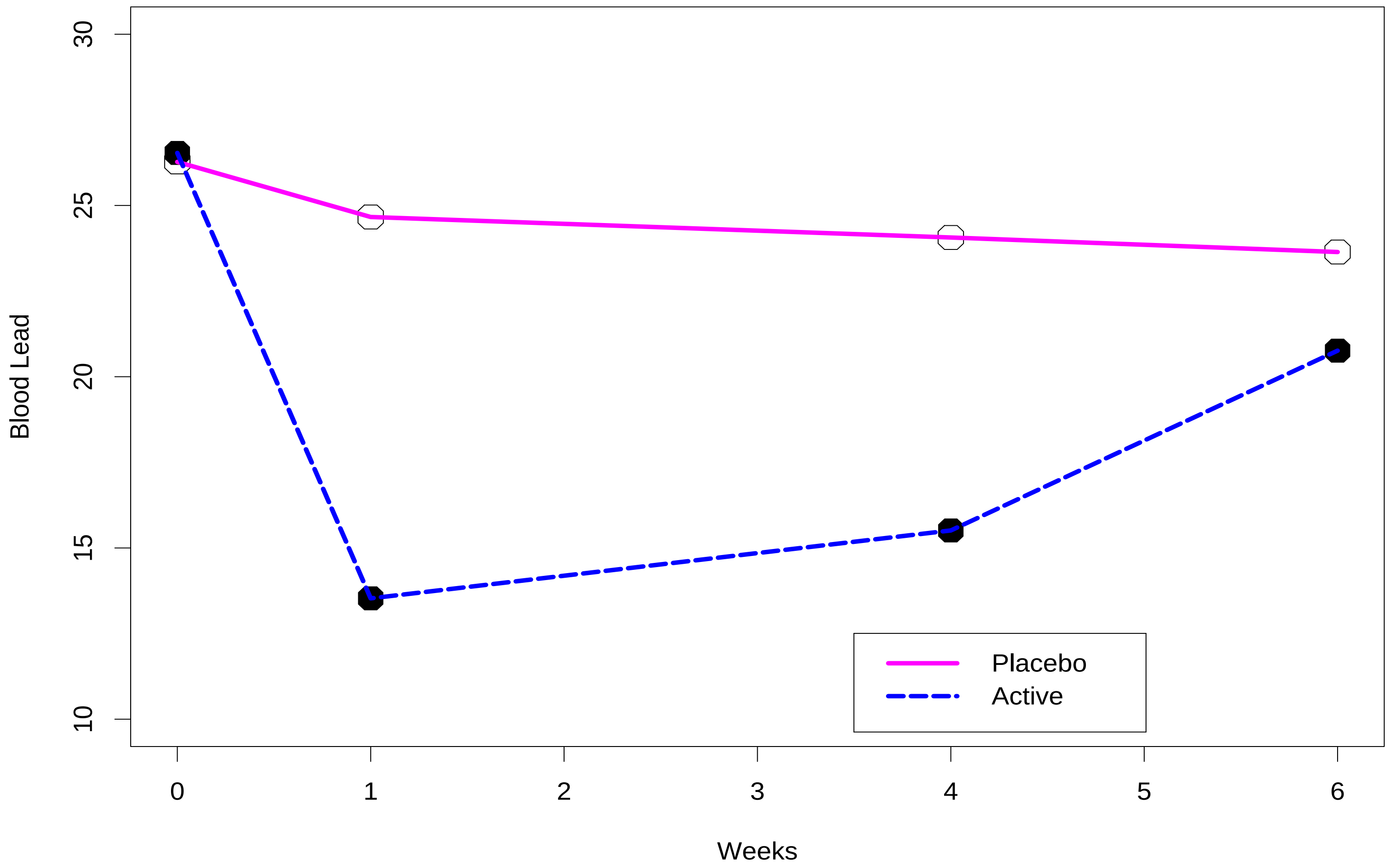
to

CORRELATION and WEIGHTING

Treatment of Lead-Exposed Children (TLC)

- $\boxed{\text{Trial:}}$ In the 1990's a placebo-controlled randomized trial of a new chelating agent, *succimer*, was conducted among children with lead levels 20-44μg/dL.
- Children received up to three 26-day courses of succimer or placebo and were followed for 3 years.
- Data set with 100 children.
- $m = 50$ placebo; $m = 50$ active.
- $\boxed{\textbf{Illustrate:}}$ naive analyses and the impact of correlation.

TLC Trial – Means



Simple (naive?) Analysis of Treatment

- $\boxed{\textbf{Post Data Only}}$ – compare the mean blood lead after baseline in the TX and control groups – using 3 measurements/person, and all 100 subjects.

  ▷ **Issue(s)** =

- $\boxed{\textbf{Pre/Post Data}}$ – compare the mean blood lead after baseline to the mean blood lead at baseline for the treatment subjects only – using 4 measurements/person, and only 50 subjects.

  ▷ **Issue(s)** =

Simple Analysis: Post Only Data

	week 0	week 1	week 4	week 6
Control		β_0	β_0	β_0
Treatment		$\beta_0 + \beta_1$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1$

Post Data Only

```
. *** Analysis using POST DATA at weeks 1, 4, and 6
.
. regress y tx if week>=1

Number of obs =      300
------------------------------------------------------------------
       y |      Coef.   Std. Err.     t    P>|t|   [95% Conf. Interval]
---------+--------------------------------------------------------
      tx |     -7.526   0.8503     -8.85   0.000      -9.199    -5.852
   _cons |     24.125   0.6012     40.12   0.000      22.942    25.308
------------------------------------------------------------------
```

```
.
. regress y tx if week>=1, cluster(id)

Number of obs =     300
Regression with robust standard errors
Number of clusters (id) = 100
------------------------------------------------------------------------
        |               Robust
      y |      Coef.   Std. Err.    t    P>|t|    [95% Conf. Interval]
--------+---------------------------------------------------------------
     tx |     -7.526   1.2287    -6.12   0.000      -9.964      -5.087
  _cons |     24.125   0.7458    32.35   0.000      22.645      25.605
------------------------------------------------------------------------
```

Simple Analysis: Pre/Post for Treatment Group Only

	week 0	week 1	week 4	week 6
Control				
Treatment	β_0	$\beta_0 + \beta_1$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1$

Pre/Post Data, TX Group Only

```
.
. *** Analysis using PRE/POST for treatment subjects
.
. regress y post if tx==1

Number of obs =      200
-------------------------------------------------------------------
       y |     Coef.   Std. Err.     t     P>|t|    [95% Conf. Interval]
---------+---------------------------------------------------------
    post |  -9.940    1.3093     -7.59    0.000     -12.522     -7.358
   _cons |  26.540    1.1339     23.41    0.000      24.303     28.776
------------------------------------------------------------------
```

```
.
. regress y post if tx==1, cluster(id)

Number of obs =     200
Regression with robust standard errors
Number of clusters (id) = 50
-----------------------------------------------------------------
        |             Robust
      y |   Coef.   Std. Err.   t    P>|t|   [95% Conf. Interval]
--------+--------------------------------------------------------
   post |  -9.940   0.8680   -11.45   0.000    -11.685   -8.196
  _cons |  26.540   0.7118    37.28   0.000     25.109   27.970
----------------------------------------------------------------
```

Dependent Data and Proper Variance Estimates

Let $X_{ij} = 0$ denote placebo assignment and $X_{ij} = 1$ denote active treatment.

(1) Consider (Y_{i1}, Y_{i2}) with $(X_{i1}, X_{i2}) = (0, 0)$ for $i = 1 : n$ and $(X_{i1}, X_{i2}) = (1, 1)$ for $i = (n + 1) : 2n$

$$
\begin{aligned}
\hat{\mu}_0 &= \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2} Y_{ij} \\
\hat{\mu}_1 &= \frac{1}{2n} \sum_{i=n+1}^{2n} \sum_{j=1}^{2} Y_{ij} \\
\text{var}(\hat{\mu}_1 - \hat{\mu}_0) &= \frac{1}{n} \{\sigma^2 (1 + \rho)\}
\end{aligned}
$$

Scenario 1

subject	control		treatment	
	time 1	time 2	time 1	time 2
ID $=$ 101	$Y_{1,1}$	$Y_{1,2}$		
ID $=$ 102	$Y_{2,1}$	$Y_{2,2}$		
ID $=$ 103	$Y_{3,1}$	$Y_{3,2}$		
ID $=$ 104			$Y_{4,1}$	$Y_{4,2}$
ID $=$ 105			$Y_{5,1}$	$Y_{5,2}$
ID $=$ 106			$Y_{6,1}$	$Y_{6,2}$

Dependent Data and Proper Variance Estimates

$\boxed{\textbf{(2)}}$ Consider (Y_{i1}, Y_{i2}) with $(X_{i1}, X_{i2}) = (0, 1)$ for $i = 1 : n$ and $(X_{i1}, X_{i2}) = (1, 0)$ for $i = (n + 1) : 2n$

$$\hat{\mu}_0 = \frac{1}{2n}\left\{\sum_{i=1}^{n} Y_{i1} + \sum_{i=n+1}^{2n} Y_{i2}\right\}$$

$$\hat{\mu}_1 = \frac{1}{2n}\left\{\sum_{i=1}^{n} Y_{i2} + \sum_{i=n+1}^{2n} Y_{i1}\right\}$$

$$\text{var}(\hat{\mu}_1 - \hat{\mu}_0) = \frac{1}{n}\{\sigma^2(1 - \rho)\}$$

Scenario 2

subject	control		treatment	
	time 1	time 2	time 1	time 2
ID $=$ 101	$Y_{1,1}$			$Y_{1,2}$
ID $=$ 102	$Y_{2,1}$			$Y_{2,2}$
ID $=$ 103	$Y_{3,1}$			$Y_{3,2}$
ID $=$ 104		$Y_{4,2}$	$Y_{4,1}$	
ID $=$ 105		$Y_{5,2}$	$Y_{5,1}$	
ID $=$ 106		$Y_{6,2}$	$Y_{6,1}$	

Dependent Data and Proper Variance Estimates

If we simply had $2n$ independent observations on treatment $(X = 1)$ and $2n$ independent observations on control then we'd obtain

$$\begin{aligned} \text{var}(\hat{\mu}_1 - \hat{\mu}_0) &= \frac{\sigma^2}{2n} + \frac{\sigma^2}{2n} \\ &= \frac{1}{n}\sigma^2 \end{aligned}$$

Q: What is the impact of <u>dependence</u> relative to the situation where all $(2n + 2n)$ observations are independent?

(1) $\Rightarrow$ positive dependence, $\rho > 0$, results in a loss of precision.

(2) $\Rightarrow$ positive dependence, $\rho > 0$, results in an improvement in precision!

Therefore:

- Dependent data impacts proper statements of precision.

- Dependent data may increase or decrease standard errors depending on the design.

Weighted Estimation

Consider the situation where subjects report both the number of attempts and the number of successes: (Y_i, N_i).

Examples:

live born (Y_i) in a litter (N_i)
condoms used (Y_i) in sexual encounters (N_i)
SAEs (Y_i) among total surgeries (N_i)

Q: How to combine these data from $i = 1 : m$ subjects to estimate a common rate (proportion) of successes?

Weighted Estimation

Proposal 1:

$$\hat{p}_1 = \sum_i Y_i / \sum_i N_i$$

Proposal 2:

$$\hat{p}_2 = \frac{1}{m} \sum_i Y_i / N_i$$

Simple Example:

$$\text{Data} \quad : \quad (1, 10) \;\; (2, 100)$$

$$\begin{aligned} \hat{p}_1 &= (2+1)/(110) \; = \; 0.030 \\ \hat{p}_2 &= \frac{1}{2}\{1/10 + 2/100\} = 0.051 \end{aligned}$$

Weighted Estimation

Note: Each of these estimators, $\hat{p}_1$, and $\hat{p}_2$, can be viewed as weighted estimators of the form:

$$\hat{p}_w = \left\{ \sum_i w_i \frac{Y_i}{N_i} \right\} / \sum_i w_i$$

We obtain $\hat{p}_1$ by letting $w_i = N_i$, corresponding to equal weight given each to binary outcome, Y_{ij}, $Y_i = \sum_{j=1}^{N_i} Y_{ij}$.

We obtain $\hat{p}_2$ by letting $w_i = 1$, corresponding to equal weight given to each subject.

Q: What's optimal?

Weighted Estimation

A: Whatever weights are closest to 1/variance of Y_i/N_i (stat theory called "Gauss-Markov").

- If subjects are perfectly homogeneous then

$$V(Y_i) = N_i p(1-p)$$

and $\hat{p}_1$ is best.

- If subjects are heterogeneous then, for example

$$V(Y_i) = N_i p(1-p)\{1 + (N_i - 1)\rho\}$$

and an estimator closer to $\hat{p}_2$ is best.

Summary: Role of Correlation

- Statistical inference must account for the dependence.
  - ▷ correlation impacts **standard errors**!
- Consideration as to the choice of weighting will depend on the variance/covariance of the response variables.
  - ▷ correlation impacts **regression estimates**!

INTRODUCTION

to

REGRESSION APPROACHES

Statistical Models

- $\boxed{\text{Regression model:}}$ **Groups**
  mean response as a function of covariates.
  "systematic variation"

- $\boxed{\text{Random effects:}}$ **Individuals**
  variation from subject-to-subject in trajectory.
  "random between-subject variation"

- $\boxed{\text{Within-subject variation:}}$ **Observations**
  variation of individual observations over time
  "random within-subject variation"

Groups: Scientific Questions as Regression

$\boxed{\star}$ Questions concerning the <u>rate of decline</u> refer to the time slope for FEV1:

$$E[\text{ FEV1 } \mid \boldsymbol{X} = \text{age, gender, f508}] \quad = \quad \beta_0(\boldsymbol{X}) + \beta_1(\boldsymbol{X}) \cdot \text{time}$$

$\boxed{\text{Time Scales}}$

- Let age0 $=$ age-at-entry, age_{i1}
- Let ageL $=$ time-since-entry, $\text{age}_{ij} - \text{age}_{i1}$

CF Regression Model

Model:

$$
\begin{aligned}
E[\texttt{FEV} \mid \boldsymbol{X}_i] \quad = \quad & \beta_0 \\
& +\beta_1 \cdot \texttt{age0} + \beta_2 \cdot \texttt{ageL} \\
& +\beta_3 \cdot \texttt{female} \\
& +\beta_4 \cdot \texttt{f508} = \texttt{1} + \beta_5 \cdot \texttt{f508} = \texttt{2} \\
& +\beta_6 \cdot \texttt{female} \cdot \texttt{ageL} \\
& +\beta_7 \cdot \texttt{f508} = \texttt{1} \cdot \texttt{ageL} + \beta_8 \cdot \texttt{f508} = \texttt{2} \cdot \texttt{ageL}
\end{aligned}
$$

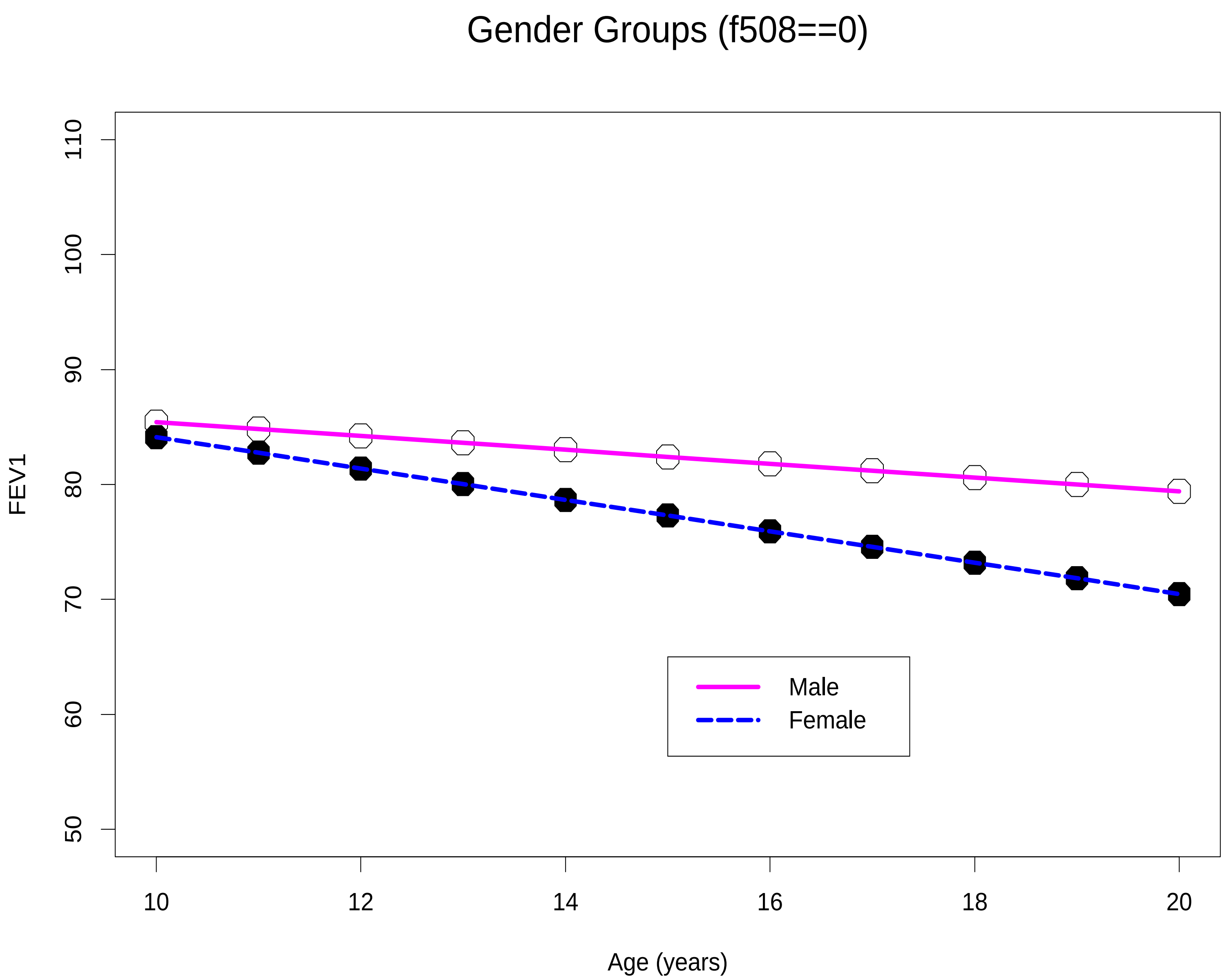
$$
= \quad \beta_0(\boldsymbol{X}_i) + \beta_1(\boldsymbol{X}_i) \cdot \texttt{ageL}
$$

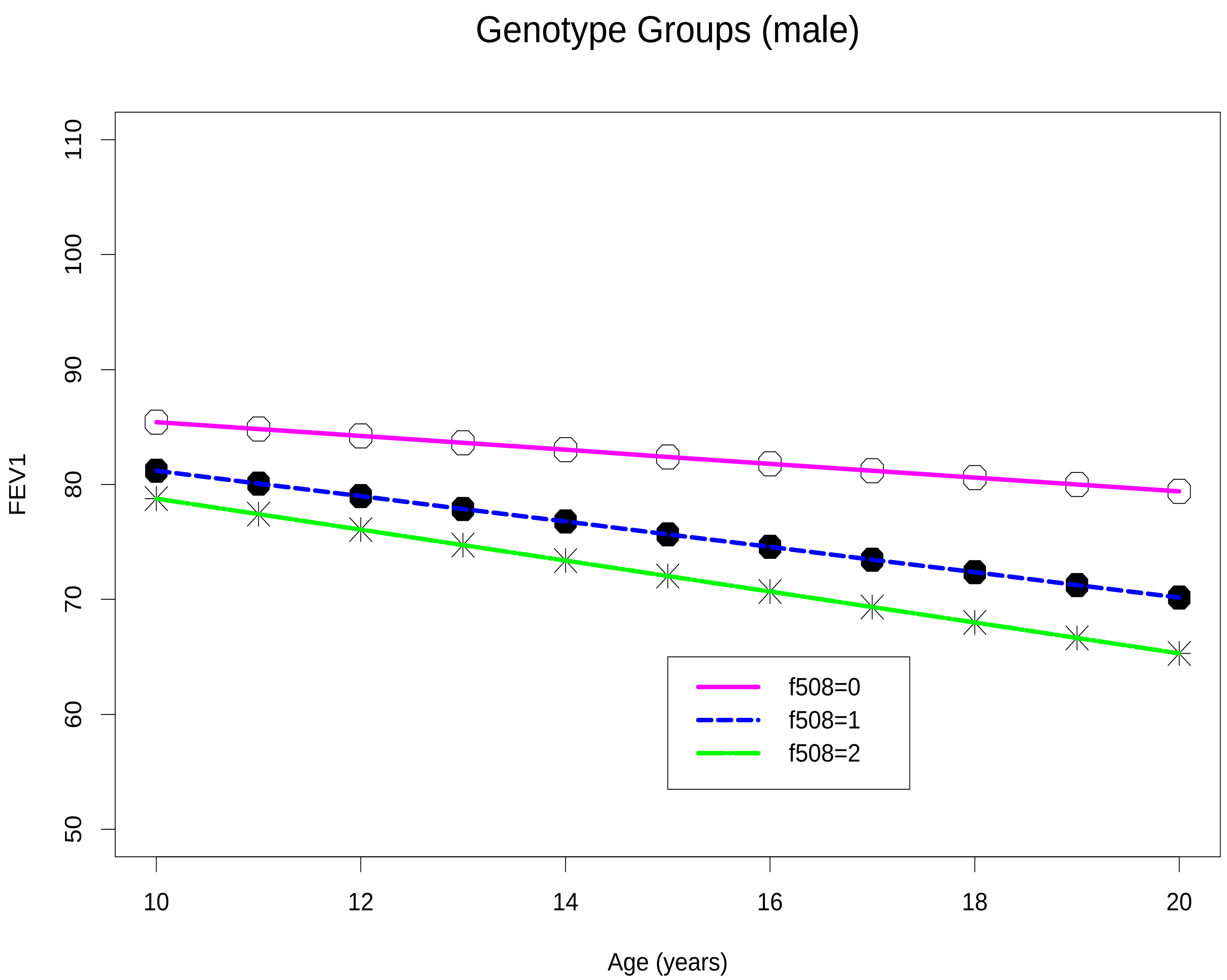
Intercept

	f508=0	f508=1	f508=2
male	$\beta_0 + \beta_1 \cdot \mathsf{age0}$	$\beta_0 + \beta_1 \cdot \mathsf{age0}$ $+\beta_4$	$\beta_0 + \beta_1 \cdot \mathsf{age0}$ $+\beta_5$
female	$\beta_0 + \beta_1 \cdot \mathsf{age0}$ $+\beta_3$	$\beta_0 + \beta_1 \cdot \mathsf{age0}$ $+\beta_3 + \beta_4$	$\beta_0 + \beta_1 \cdot \mathsf{age0}$ $+\beta_3 + \beta_5$

Slope

	f508=0	f508=1	f508=2
male	β_2	$\beta_2 + \beta_7$	$\beta_2 + \beta_8$
female	β_2 $+\beta_6$	$\beta_2 + \beta_7$ $+\beta_6$	$\beta_2 + \beta_8$ $+\beta_6$


Gender Groups (f508==0)
FEV1
110
100
90
80
70
60
50
10
12
14
16
18
20
Age (years)
Male
Female


Genotype Groups (male)
FEV1
110
100
90
80
70
60
50
10
12
14
16
18
20
Age (years)
f508=0
f508=1
f508=2

$\boxed{\text{Define}}$

$$
\begin{aligned}
Y_{ij} &= \text{FEV1 for subject } i \text{ at time } t_{ij} \\
\boldsymbol{X}_i &= (\boldsymbol{X}_{ij}, \ldots, \boldsymbol{X}_{in_i}) \\
\boldsymbol{X}_{ij} &= (X_{ij,1}, X_{ij,2}, \ldots, X_{ij,p}) \\
& \quad \text{age0, ageL, gender, genotype}
\end{aligned}
$$

Issue: Response variables measured on the same subject are correlated.

$$\text{cov}(Y_{ij}, Y_{ik}) \neq 0$$

Some Notation

- It is useful to have some notation that can be used to discuss the stack of data that correspond to each subject.
- Let n_i denote the number of observations for subject i.
- $\boxed{\text{Define:}}$

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}$$

- If the subjects are observed at a common set of times $t_1, t_2, \ldots, t_m$ then $E(Y_{ij}) = \mu_j$ denotes the mean of the population at time t_j.

Dependence and Correlation

- Recall that observations are termed **independent** when deviation in one variable does not predict deviation in the other variable.
  - ▷ Given two subjects with the same age and gender, then the blood pressure for patient ID=212 is <u>not</u> predictive of the blood pressure for patient ID=334.
- Observations are called **dependent** or **correlated** when one variable does predict the value of another variable.
  - ▷ The LDL cholesterol of patient ID=212 at age 57 <u>is</u> predictive of the LDL cholesterol of patient ID=212 at age 60.

Dependence and Correlation

- $\boxed{\text{Recall:}}$ The variance of a variable, Y_{ij} (fix time t_j for now) is defined as:

$$\begin{aligned} \sigma_j^2 &= E\left[(Y_{ij} - \mu_j)^2\right] \\ &= E\left[(Y_{ij} - \mu_j)(Y_{ij} - \mu_j)\right] \end{aligned}$$

- The variance measures the average distance that an observation falls away from the mean.

Dependence and Correlation

- $\boxed{\text{Define:}}$ The **covariance** of two variables, Y_{ij}, and Y_{ik} (fix t_j and t_k) is defined as:

$$\sigma_{jk} = E\left[(Y_{ij} - \mu_j)(Y_{ik} - \mu_k)\right]$$

- The covariance measures whether, on average, departures in one variable, $Y_{ij} - \mu_j$, "go together with" departures in a second variable, $Y_{ik} - \mu_k$.

- In simple linear regression of Y_{ij} on Y_{ik} the regression coefficient β_1 in $E(Y_{ij} \mid Y_{ik}) = \beta_0 + \beta_1 \cdot Y_{ik}$ is the covariance divided by the variance of Y_{ik}:

$$\beta_1 = \frac{\sigma_{jk}}{\sigma_k^2}$$

Dependence and Correlation

- $\boxed{\text{Define:}}$ The **correlation** of two variables, Y_{ij}, and Y_{ik} (fix t_j and t_k) is defined as:

$$\rho_{jk} = \frac{E\left[(Y_{ij} - \mu_j)(Y_{ik} - \mu_k)\right]}{\sigma_j \sigma_k}$$

- The correlation is a measure of dependence that takes values between -1 and $+1$.
- Recall that a correlation of 0.0 implies that the two measures are unrelated (linearly).
- Recall that a correlation of 1.0 implies that the two measures fall perfectly on a line – one exactly predicts the other!

Why interest in covariance and/or correlation?

- Recall that on earlier pages our standard error for the sample mean difference $\widehat{\mu}_1 - \widehat{\mu}_0$ depends on ρ.
- In general a statistical model for the outcomes $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})$ requires the following:
  - ▷ **Means**: μ_j
  - ▷ **Variances**: σ_j^2
  - ▷ **Covariances**: σ_{jk}, or correlations ρ_{jk}.
- Therefore, one approach to making inferences based on longitudinal data is to construct a model for each of these three components.

Something new to model...

$$
\begin{aligned}
\mathsf{cov}(Y_i) &= \begin{bmatrix} \mathsf{var}(Y_{i1}) & \mathsf{cov}(Y_{i1}, Y_{i2}) & \dots & \mathsf{cov}(Y_{i1}, Y_{in_i}) \\ \mathsf{cov}(Y_{i2}, Y_{i1}) & \mathsf{var}(Y_{i2}) & \dots & \mathsf{cov}(Y_{i2}, Y_{in_i}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{cov}(Y_{in_i}, Y_{i1}) & \mathsf{cov}(Y_{in_i}, Y_{i2}) & \dots & \mathsf{var}(Y_{in_i}) \end{bmatrix} \\
&= \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \dots & \sigma_1\sigma_{n_i}\rho_{1n_i} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2^2 & \dots & \sigma_2\sigma_{n_i}\rho_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_i}\sigma_1\rho_{n_i1} & \sigma_{n_i}\sigma_2\rho_{n_i2} & \dots & \sigma_{n_i}^2 \end{bmatrix}
\end{aligned}
$$

TLC Trial – Covariances

Placebo

	y0	y1	y4	y6
y0	25.2	22.7	24.3	21.4
y1	22.7	29.8	27.0	23.4
y4	24.3	27.0	33.1	28.2
y6	21.4	23.4	28.2	31.8

Active

	y0	y1	y4	y6
y0	25.2	15.5	15.1	23.0
y1	15.5	58.9	44.0	36.0
y4	15.1	44.0	61.7	33.0
y6	23.0	36.0	33.0	85.5

TLC Trial – Correlations

Placebo

	y0	y1	y4	y6
y0	1.00	0.83	0.84	0.76
y1	0.83	1.00	0.86	0.76
y4	0.84	0.86	1.00	0.87
y6	0.76	0.76	0.87	1.00

Active

	y0	y1	y4	y6
y0	1.00	0.40	0.38	0.50
y1	0.40	1.00	0.73	0.51
y4	0.38	0.73	1.00	0.45
y6	0.50	0.51	0.45	1.00

Mean and Covariance Models for FEV1

$\boxed{\text{Models:}}$

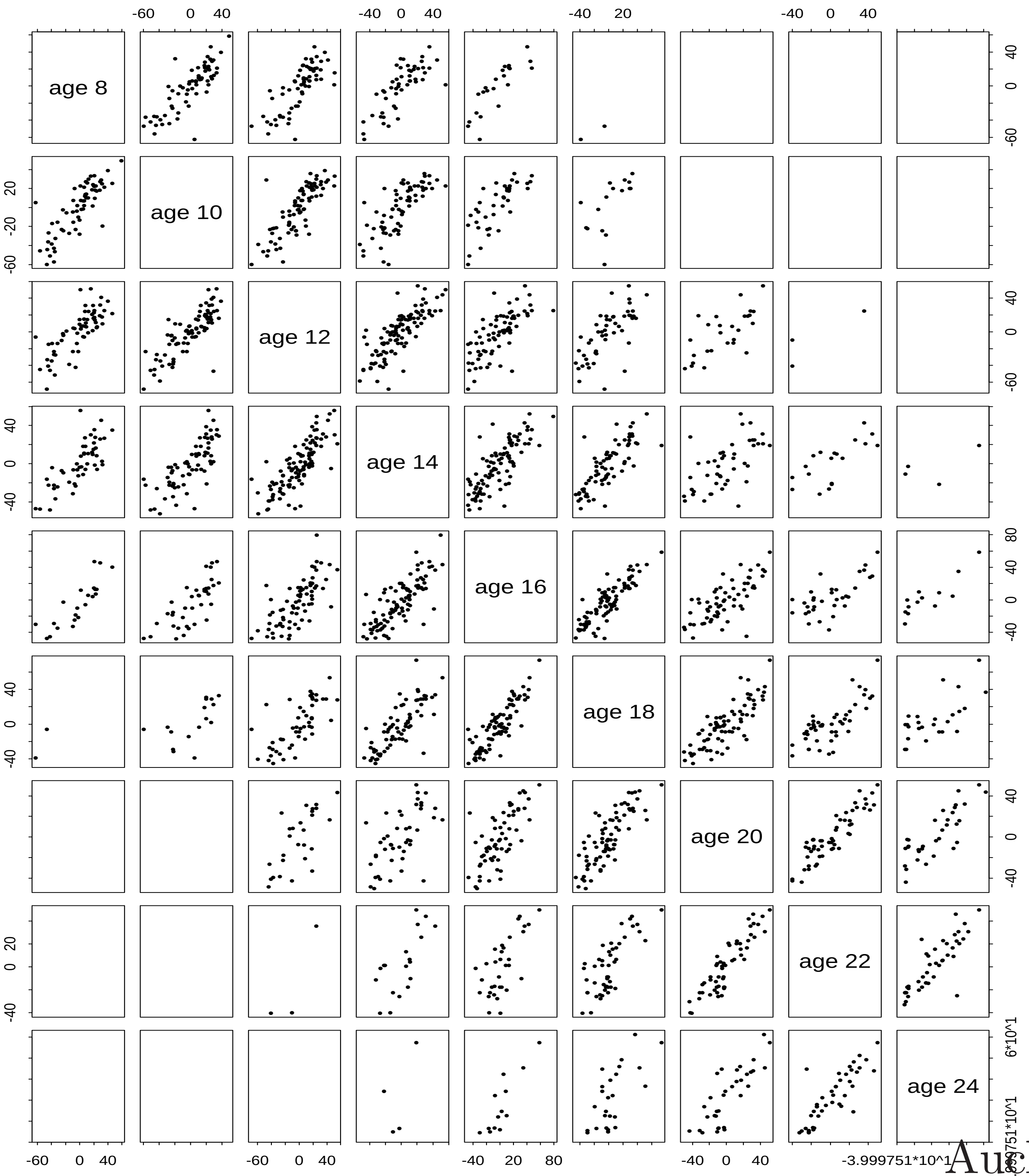
$$E(Y_{ij} \mid \boldsymbol{X}_i) \quad = \quad \mu_{ij} \ \ (\texttt{regression}) \ \textbf{Groups}$$

$$\text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) \quad = \quad \boldsymbol{\Sigma}_i = \underbrace{\textbf{between-subjects}}_{\texttt{individual-to-individual}} + \underbrace{\textbf{within-subjects}}_{\texttt{observation-to-observation}}$$

Q: What are appropriate covariance models for the FEV1 data?

Individual-to-Individual variation?

Observation-to-Observation variation?


age 8
age 10
age 12
age 14
age 16
age 18
age 20
age 22
age 24
-3.999751*10^1

How to build models for correlation?

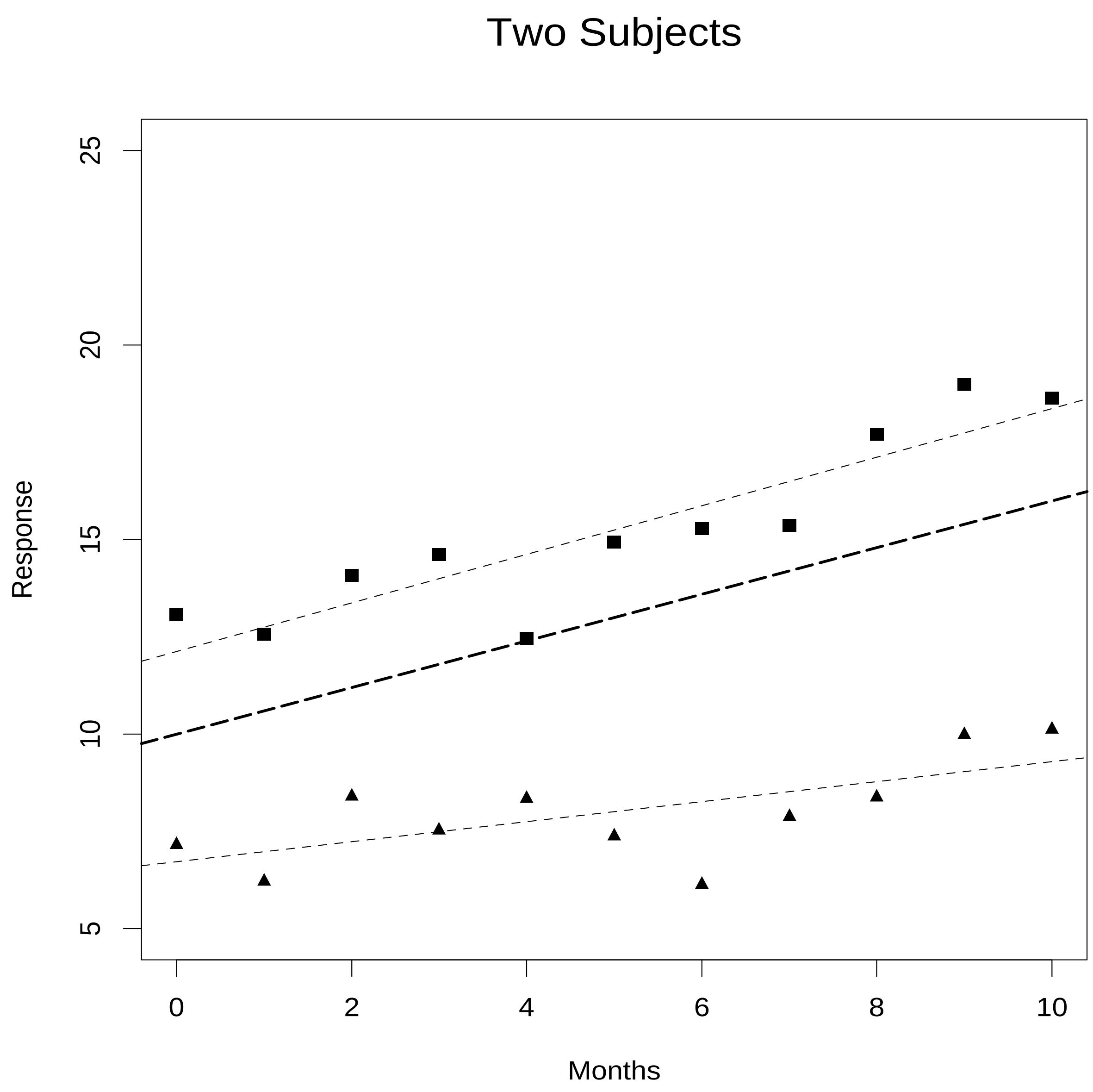
- $\boxed{\text{Mixed models}}$
  - "random effects"
  - between-subject variability
  - within-subject similarity due to sharing trajectory
- $\boxed{\text{Serial correlation}}$
  - close in time implies strong similarity
  - correlation decreases as time separation increases

Toward the Linear Mixed Model

- $\boxed{\text{Regression model:}}$
  mean response as a function of covariates.
  "systematic variation"

- $\boxed{\text{Random effects:}}$ **Individuals**
  variation from subject-to-subject in trajectory.
  "random between-subject variation"

- $\boxed{\text{Within-subject variation:}}$
  variation of individual observations over time
  "random within-subject variation"


Two Subjects
Response
Months
25
20
15
10
5
0
2
4
6
8
10

Levels of Analysis

- We first consider the distribution of **measurements** within **subjects**:

$$
\begin{aligned}
Y_{ij} &= \beta_{0,i} + \beta_{1,i} \cdot t_{ij} + e_{ij} \\
e_{ij} &\sim \mathcal{N}(0, \sigma^2) \\
E[\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{\beta}_i] &= \beta_{0,i} + \beta_{1,i} \cdot t_{ij} \\
&= [1, \texttt{time}_{ij}] \begin{bmatrix} \beta_{0,i} \\ \beta_{1,i} \end{bmatrix} \\
&= \boldsymbol{X}_i \boldsymbol{\beta}_i
\end{aligned}
$$

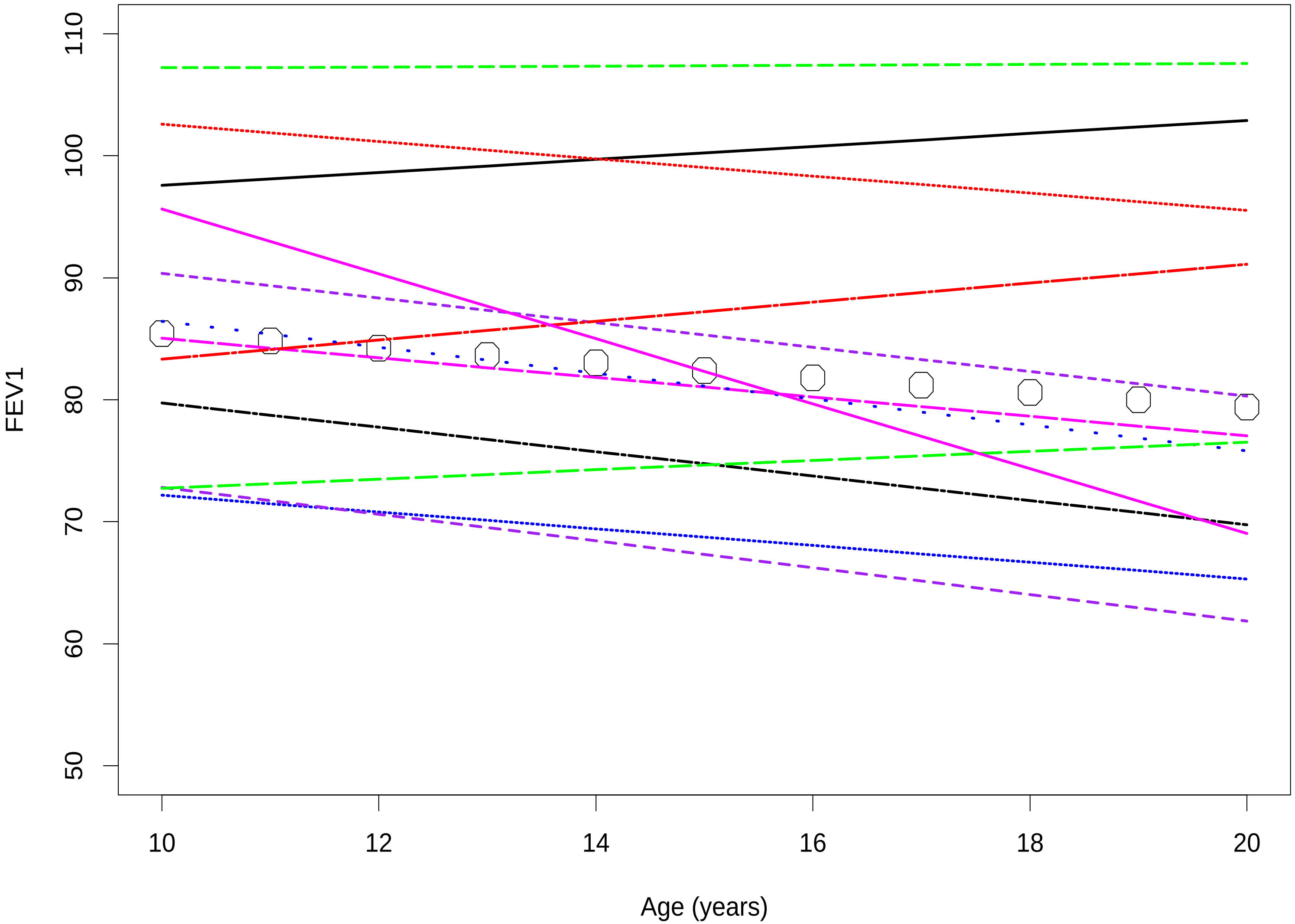
Levels of Analysis

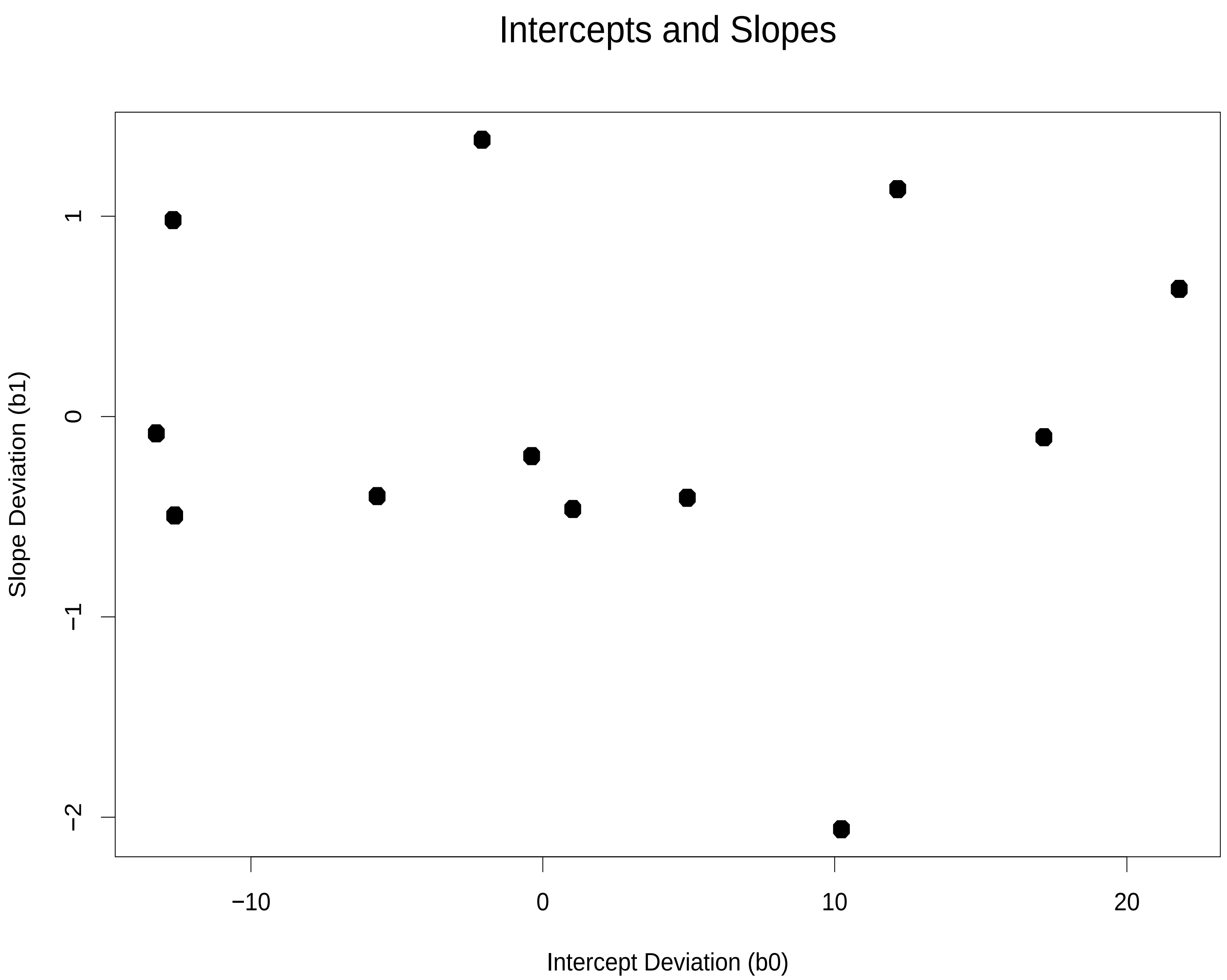
- We can equivalently separate the subject-specific regression coefficients into the **average coefficient** and the **specific departure** for subject i:
  - ▷ $\beta_{0,i} = \beta_0 + b_{0,i}$
  - ▷ $\beta_{1,i} = \beta_1 + b_{1,i}$
- This allows another perspective:

$$\begin{aligned} Y_{ij} &= \beta_{0,i} + \beta_{1,i} \cdot t_{ij} + e_{ij} \\ &= (\beta_0 + \beta_1 \cdot t_{ij}) + (b_{0,i} + b_{1,i} \cdot t_{ij}) + e_{ij} \end{aligned}$$

$$E[\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{\beta}_i] = \underbrace{\boldsymbol{X}_i \boldsymbol{\beta}}_{\textbf{mean model}} + \underbrace{\boldsymbol{X}_i \boldsymbol{b}_i}_{\textbf{between-subject}}$$

Sample of Lines




Intercepts and Slopes
Slope Deviation (b1)
Intercept Deviation (b0)
−10
0
10
20
−2
−1
0
1

Levels of Analysis

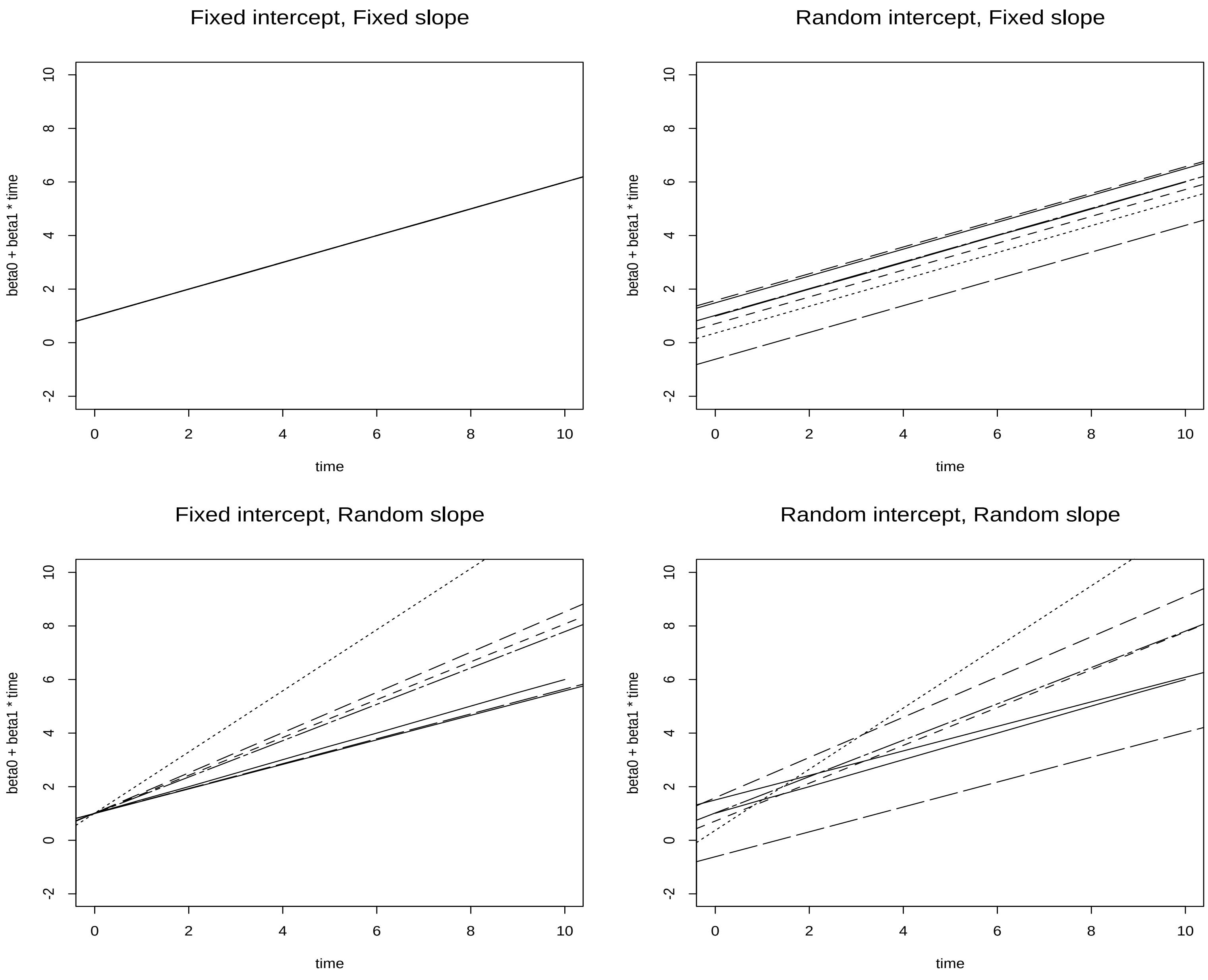
- Next we consider the distribution of **patterns (parameters)** among subjects:

$$\boldsymbol{\beta}_i \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{D})$$

equivalently

$$\boldsymbol{b}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{D})$$

$$\star\star\star \quad \boldsymbol{Y}_i = \underbrace{\boldsymbol{X}_i\boldsymbol{\beta}}_{\textbf{mean model}} + \underbrace{\boldsymbol{X}_i\boldsymbol{b}_i}_{\textbf{between-subject}} + \underbrace{\boldsymbol{e}_i}_{\textbf{within-subject}}$$


Fixed intercept, Fixed slope
Random intercept, Fixed slope
Fixed intercept, Random slope
Random intercept, Random slope
beta0 + beta1 * time
time
-2
0
2
4
6
8
10

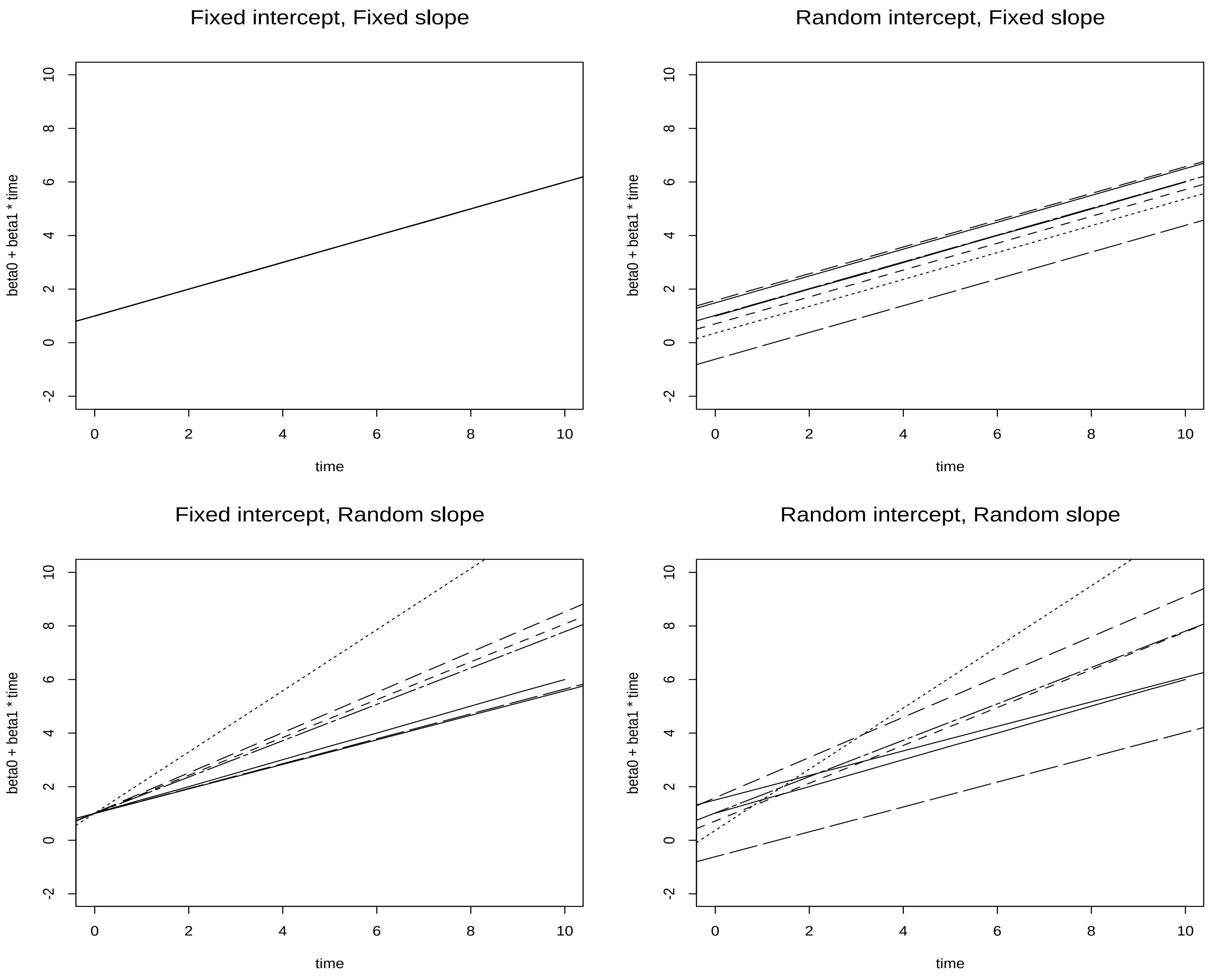
Between-subject Variation

- We can use the idea of random effects to allow different types of between-subject heterogeneity:
- The magnitude of heterogeneity is characterized by $\boldsymbol{D}$:

$$\begin{aligned}
\boldsymbol{b}_i &= \begin{bmatrix} b_{0,i} \\ b_{1,i} \end{bmatrix} \\
\mathsf{var}(\boldsymbol{b}_i) &= \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}
\end{aligned}$$

Between-subject Variation

- The components of $\boldsymbol{D}$ can be interpreted as:
  - ▷ $\sqrt{D_{11}}$ – the typical subject-to-subject deviation in the overall **level** of the response.
  - ▷ $\sqrt{D_{22}}$ – the typical subject-to-subject deviation in the **change** (time slope) of the response.
  - ▷ D_{12} – the covariance between individual intercepts and slopes.
    - ∗ If positive then subjects with **high levels** also have **high rates** of change.
    - ∗ If negative then subjects with **high levels** have **low rates** of change.


Fixed intercept, Fixed slope
Random intercept, Fixed slope
Fixed intercept, Random slope
Random intercept, Random slope
beta0 + beta1 * time
time
-2
0
2
4
6
8
10

Between-subject Variation: Examples

- $\boxed{\text{No random effects:}}$

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 \cdot t_{ij} + e_{ij} \\ &= [1, \texttt{time}_{ij}]\boldsymbol{\beta} + e_{ij} \end{aligned}$$

- $\boxed{\text{Random intercepts:}}$

$$\begin{aligned} Y_{ij} &= (\beta_0 + \beta_1 \cdot t_{ij}) + b_{0,i} + e_{ij} \\ &= [1, \texttt{time}_{ij}]\boldsymbol{\beta} + [\ 1\ ]b_{0,i} + e_{ij} \end{aligned}$$

- $\boxed{\text{Random intercepts and slopes:}}$

$$\begin{aligned} Y_{ij} &= (\beta_0 + \beta_1 \cdot t_{ij}) + b_{0,i} + b_{1,i} \cdot t_{ij} + e_{ij} \\ &= [1, \texttt{time}_{ij}]\boldsymbol{\beta} + [1, \texttt{time}_{ij}]\boldsymbol{b}_i + e_{ij} \end{aligned}$$

Toward the Linear Mixed Model

- $\boxed{\text{Regression model:}}$
  mean response as a function of covariates.
  "systematic variation"

- $\boxed{\text{Random effects:}}$
  variation from subject-to-subject in trajectory.
  "random between-subject variation"

- $\boxed{\text{Within-subject variation:}}$ **Observation**
  variation of individual observations over time
  "random within-subject variation"

Covariance Models

Serial Models

- Linear mixed models assume that each subject follows his/her own line. In some situations the dependence is more **local** meaning that observations close in time are more similar than those far apart in time.

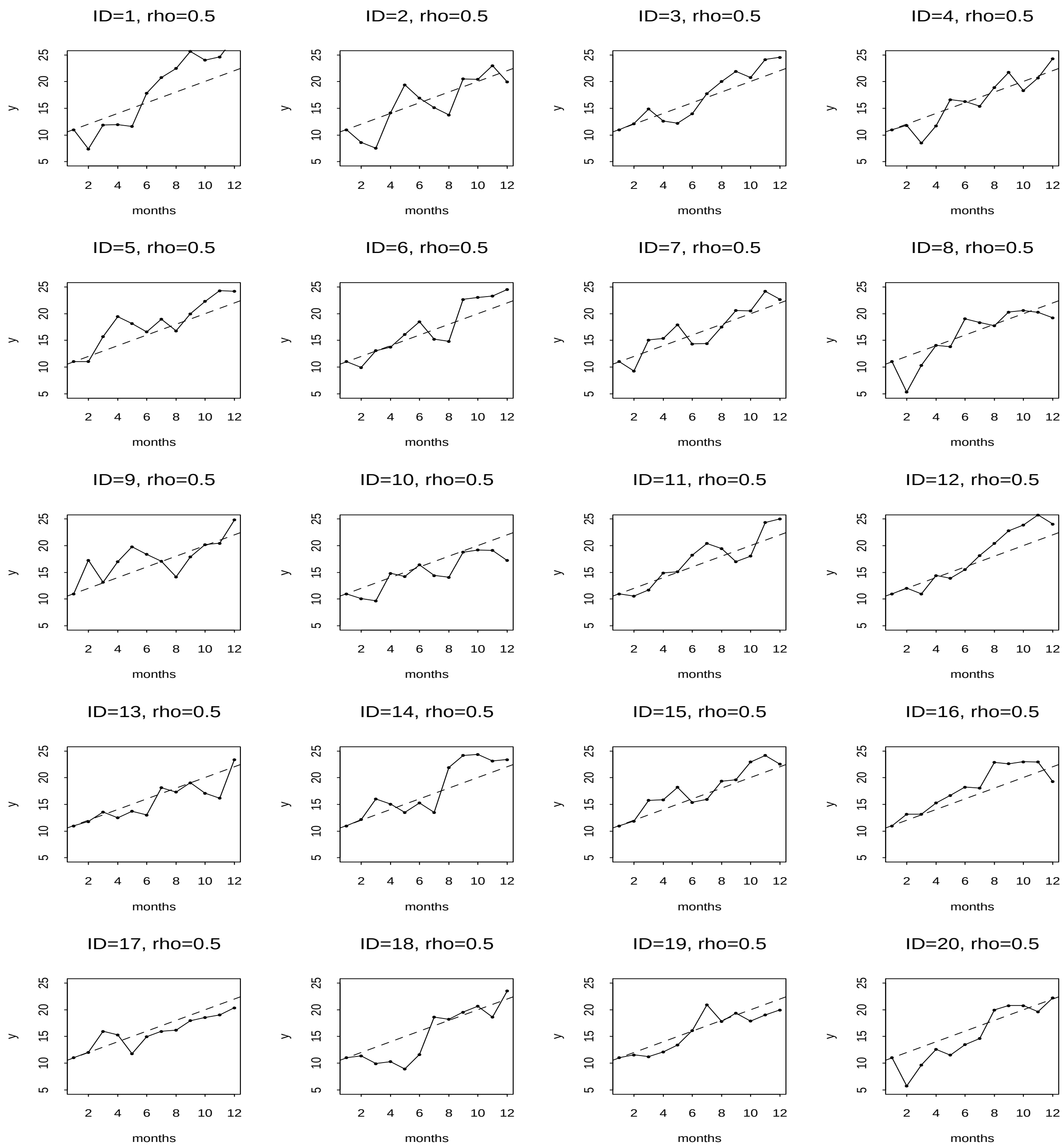
Covariance Models

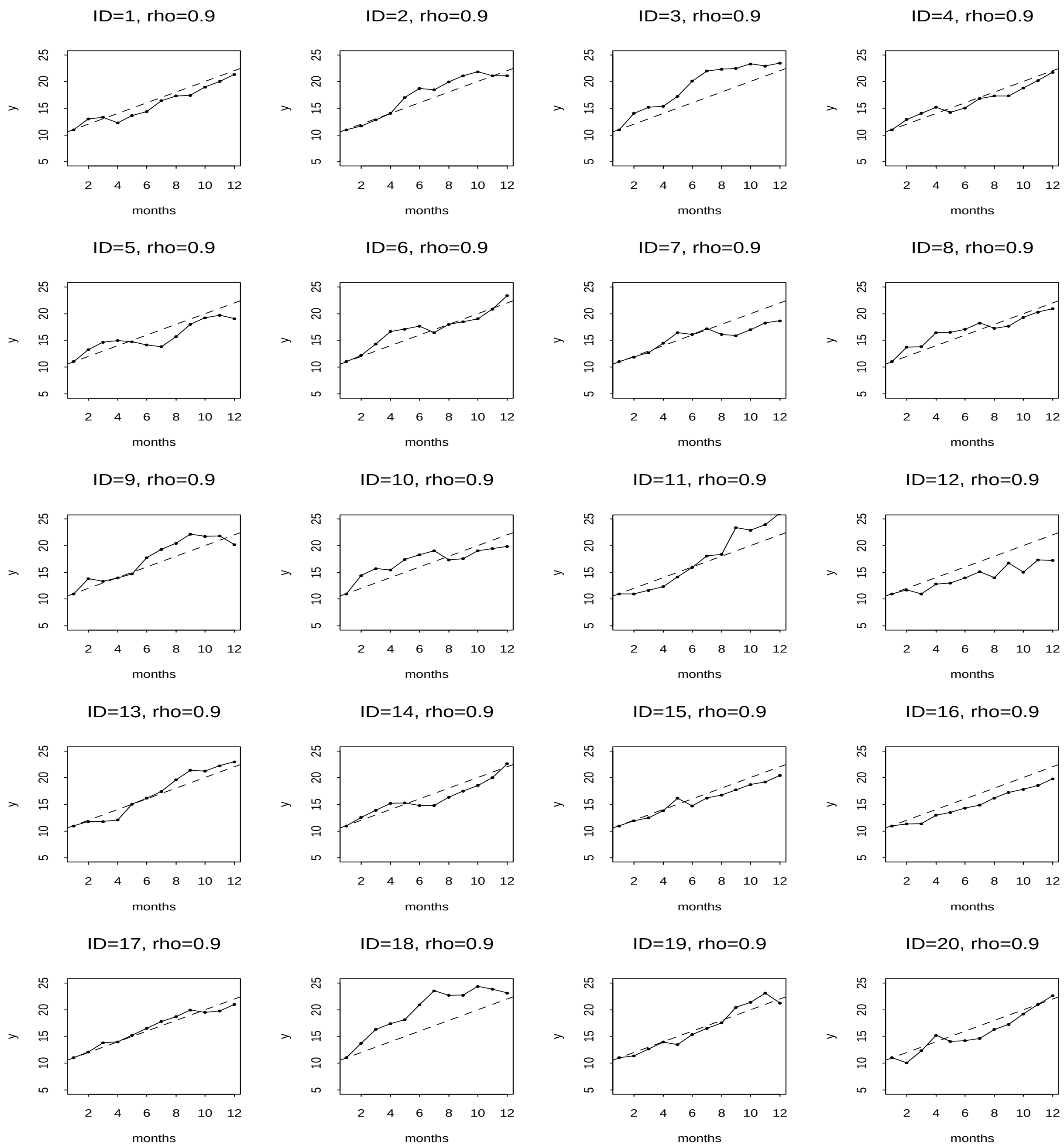
$\boxed{\text{Define}}$

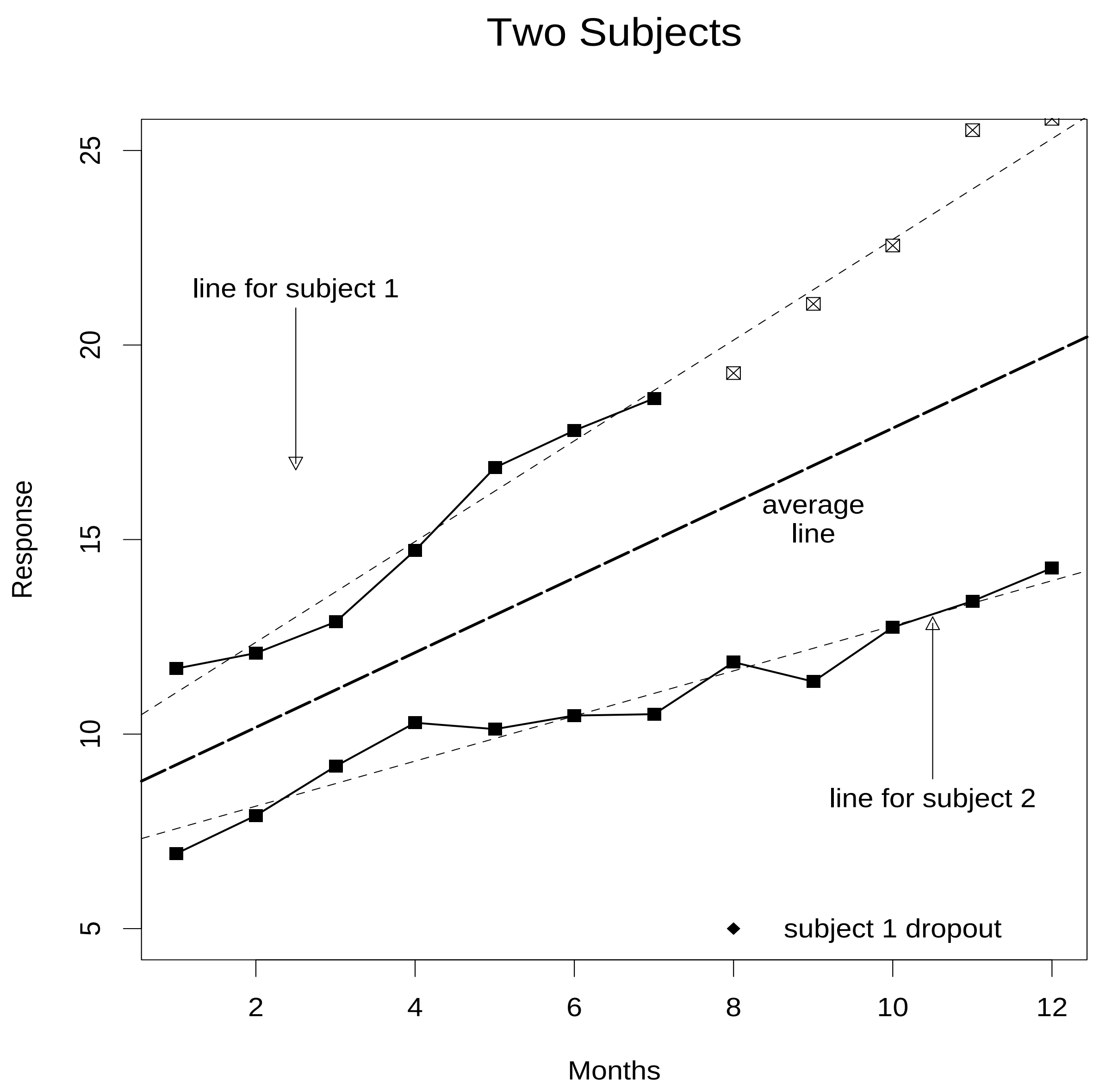
$$
\begin{aligned}
e_{ij} &= \rho \cdot e_{ij-1} + \epsilon_{ij} \\
\\
\epsilon_{ij} &\sim \mathcal{N}(0, \sigma^2(1-\rho^2)\ ) \\
\epsilon_{i0} &\sim \mathcal{N}(0, \sigma^2)
\end{aligned}
$$

This leads to **autocorrelated** errors:

$$\mathsf{cov}(e_{ij}, e_{ik}) = \sigma^2 \rho^{|j-k|}$$


ID=1, rho=0.5
ID=2, rho=0.5
ID=3, rho=0.5
ID=4, rho=0.5
ID=5, rho=0.5
ID=6, rho=0.5
ID=7, rho=0.5
ID=8, rho=0.5
ID=9, rho=0.5
ID=10, rho=0.5
ID=11, rho=0.5
ID=12, rho=0.5
ID=13, rho=0.5
ID=14, rho=0.5
ID=15, rho=0.5
ID=16, rho=0.5
ID=17, rho=0.5
ID=18, rho=0.5
ID=19, rho=0.5
ID=20, rho=0.5
y
5 10 15 20 25
2 4 6 8 10 12
months


ID=1, rho=0.9
ID=2, rho=0.9
ID=3, rho=0.9
ID=4, rho=0.9
ID=5, rho=0.9
ID=6, rho=0.9
ID=7, rho=0.9
ID=8, rho=0.9
ID=9, rho=0.9
ID=10, rho=0.9
ID=11, rho=0.9
ID=12, rho=0.9
ID=13, rho=0.9
ID=14, rho=0.9
ID=15, rho=0.9
ID=16, rho=0.9
ID=17, rho=0.9
ID=18, rho=0.9
ID=19, rho=0.9
ID=20, rho=0.9
y
months
5
10
15
20
25
2
4
6
8
10
12


Two Subjects
Response
Months
line for subject 1
average
line
line for subject 2
subject 1 dropout
5
10
15
20
25
2
4
6
8
10
12

Toward the Linear Mixed Model

- $\boxed{\text{Regression model:}}$
  mean response as a function of covariates.
  "systematic variation"

- $\boxed{\text{Random effects:}}$
  variation from subject-to-subject in trajectory.
  "random between-subject variation"

- $\boxed{\text{Within-subject variation:}}$
  variation of individual observations over time
  "random within-subject variation"

Session Three Summary

- Role of correlation
  - ▷ Impact proper standard errors
  - ▷ Used to weight individuals (clusters)

- Models for correlation / covariance
  - ▷ Regression: **Group**-to-Group variation
  - ▷ Random effects: **Individual**-to-Individual variation
  - ▷ Serial correlation: **Observation**-to-Observation variation

EXTRA: Mixed Models and Covariances/Correlation

- **Q**: What is the correlation between outcomes Y_{ij} and Y_{ik} under these random effects models?
- $\boxed{\text{Random Intercept Model}}$

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 t_{ij} + b_{0,i} + e_{ij} \\ Y_{ik} &= \beta_0 + \beta_1 t_{ik} + b_{0,i} + e_{ik} \end{aligned}$$

$$\begin{aligned} \mathrm{var}(Y_{ij}) &= \mathrm{var}(b_{0,i}) + \mathrm{var}(e_{ij}) \\ &= D_{11} + \sigma^2 \end{aligned}$$

$$\begin{aligned} \mathrm{cov}(Y_{ij}, Y_{ik}) &= \mathrm{cov}(b_{0,i} + e_{ij}, b_{0,i} + e_{ik}) \\ &= D_{11} \end{aligned}$$

EXTRA: Mixed Models and Covariances/Correlation

- $\boxed{\text{Random Intercept Model}}$

$$\begin{aligned}\text{corr}(Y_{ij}, Y_{ik}) &= \frac{D_{11}}{\sqrt{D_{11}+\sigma^2}\sqrt{D_{11}+\sigma^2}} \\ &= \frac{D_{11}}{D_{11}+\sigma^2} = \frac{\textbf{between var}}{\textbf{between var} + \textbf{within var}}\end{aligned}$$

- Therefore, any two outcomes have the same correlation. Doesn't depend on the specific times, nor on the distance between the measurements.

- "**Exchangeable**" correlation model.

- Assuming: var$(e_{ij}) = \sigma^2$, and cov$(e_{ij}, e_{ik}) = 0$.

EXTRA: Mixed Models and Covariances/Correlation

- $\boxed{\text{Random Intercept and Slope Model}}$

$$
\begin{aligned}
Y_{ij} &= (\beta_0 + \beta_1 t_{ij}) + (b_{0,i} + b_{1,i} t_{ij}) + e_{ij} \\
Y_{ik} &= (\beta_0 + \beta_1 t_{ik}) + (b_{0,i} + b_{1,i} t_{ik}) + e_{ik}
\end{aligned}
$$

$$
\begin{aligned}
\mathsf{var}(Y_{ij}) &= \mathsf{var}(b_{0,i} + b_{1,i} t_{ij}) + \mathsf{var}(e_{ij}) \\
&= D_{11} + 2 \cdot D_{12} t_{ij} + D_{22} t_{ij}^2 + \sigma^2
\end{aligned}
$$

$$
\begin{aligned}
\mathsf{cov}(Y_{ij}, Y_{ik}) &= \mathsf{cov}[(b_{0,i} + b_{1,i} t_{ij} + e_{ij}), (b_{0,i} + b_{1,i} t_{ik} + e_{ik})] \\
&= D_{11} + D_{12}(t_{ij} + t_{ik}) + D_{22} t_{ij} t_{ik}
\end{aligned}
$$

EXTRA: Mixed Models and Covariances/Correlation

- $\boxed{\text{Random Intercept and Slope Model}}$

$$\begin{aligned}\rho_{ijk} &= \text{corr}(Y_{ij}, Y_{ik}) \\ &= \frac{D_{11} + D_{12}(t_{ij} + t_{ik}) + D_{22} t_{ij} t_{ik}}{\sqrt{D_{11} + 2 \cdot D_{12} t_{ij} + D_{22} t_{ij}^2 + \sigma^2}\sqrt{D_{11} + 2 \cdot D_{12} t_{ik} + D_{22} t_{ik}^2 + \sigma^2}}\end{aligned}$$

- Therefore, two outcomes may not have the same correlation. Correlation depends on the specific times for the observations, and does not have a simple form.

- Assuming: $\text{var}(e_{ij}) = \sigma^2$, and $\text{cov}(e_{ij}, e_{ik}) = 0$.