Analysis of Correlated Data

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

Course Outline

- Examples of longitudinal data
- Correlation and weighting
- Exploratory data analysis
  - between- and within-person variation
  - correlation / covariance
- Regression analysis
  - Models for the mean
  - Models for the correlation
  - Linear mixed model
  - GEE

Course Outline

- Missing data
- Binary and Count Data
  - Models for the mean
  - Models for the correlation
  - Generalized linear mixed model
  - GEE
  - Transition models
- Time-dependent covariates
- Multilevel models
- Design considerations

Bio & Notes

- **Patrick J. Heagerty**
  - Professor, University of Washington
  - Collaborative roles = RWJ, VA ERIC, NIAMS MCRC, K12
  - Books:

    Diggle, Heagerty, Liang & Zeger
    "Analysis of Longitudinal Data" Oxford, 2002.

    van Belle, Fisher, Heagerty & Lumley
    "Biostatistics" Wiley, 2004.
    (introductory chapter on LDA)

- **Course Notes & Slides**
  - UW Biostat 571 = Ph.D. applied core sequence
    Winter 1999, 2000, 2001, 2002
  - UM Epi 766 = Longitudinal Data Analysis / Epi
    Summer 2000 (Summer 2004 with VA/UW Biostat/Epi)
  - Second Seattle Symposium (with S. Zeger)
    Fall 2000
  - RAND short course
    Fall 2002
  - NICHD short course
    Fall 2003

Longitudinal Data Analysis

INTRODUCTION

to

EXAMPLES AND ISSUES

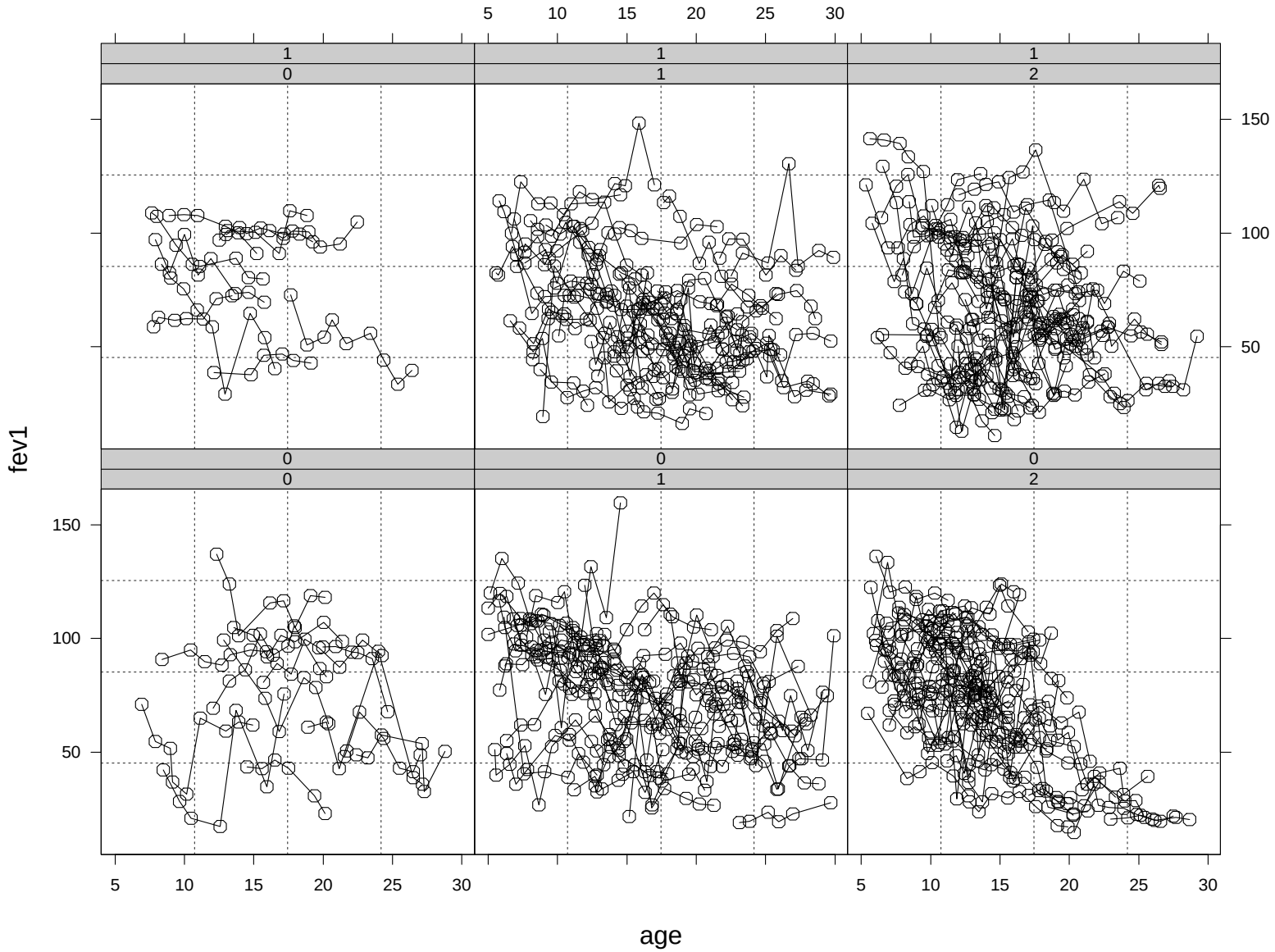
Outline

- Examples of longitudinal data
  - between- and within-person variation
  - correlation / covariance
- Scientific motivation
  - Opportunities
  - Issues
- Time scales
  - Cross-sectional contrasts
  - Longitudinal contrasts
- Correlation and weighting
  - Impact of correlation
  - Weighted estimation

Continuous Longitudinal Data

Example 1: Cystic Fibrosis and Lung Function

- There is a large registry of cystic fibrosis patient data. Annual measurements include standard pulmonary function measures: FVC, FEV1.
- primary outcome: FEV1 percent predicted.
- covariates: age, gender, genotype.
- **Q**: Does change in lung function differ by gender and/or genotype?

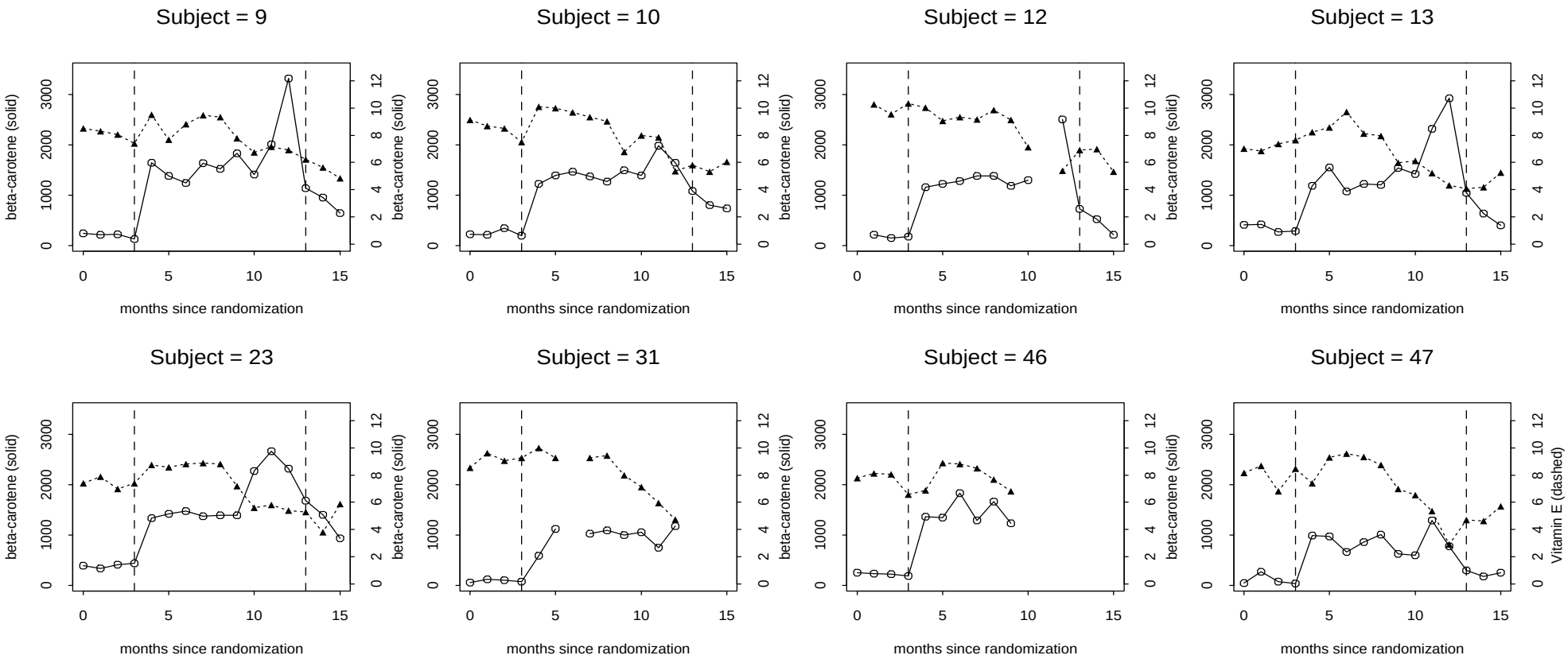

1
0
1
1
1
2
0
0
0
1
0
2
fev1
age
5
10
15
20
25
30
50
100
150

Continuous Longitudinal Data

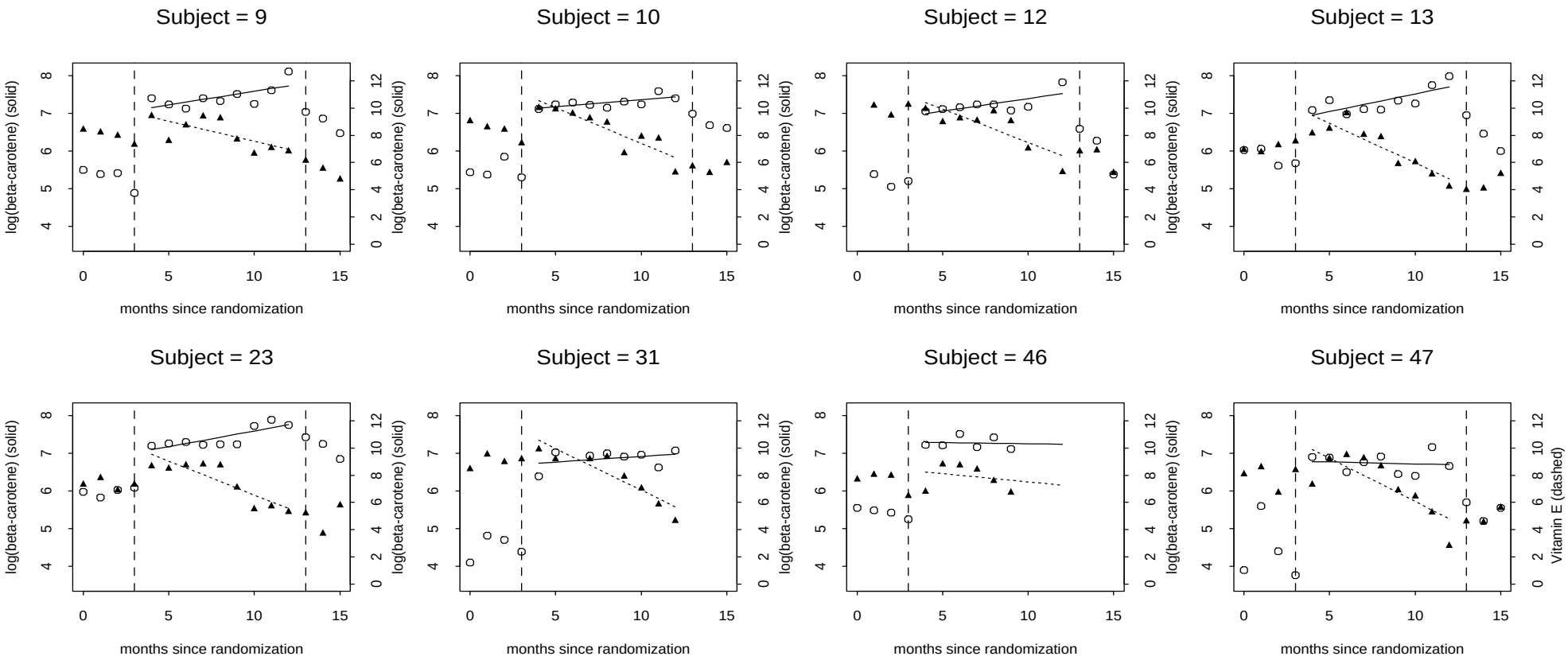
Example 2: Beta-carotene and vitamin E

- a phase II study to ascertain the pharmacokinetics of beta-carotene supplementation and the subsequent impact on vitamin E levels.
- primary outcome: plasma measures taken monthly for 3 months prior to, 9 months during, and 3 months after supplementation.
- covariates: dose (0, 15, 30, 45, 60 mg/day) and time
- **Q**: What is the time course? Dose-response? Relationship between beta-carotene and vitamin E?

Dose = 45


Subject = 9
Subject = 10
Subject = 12
Subject = 13
Subject = 23
Subject = 31
Subject = 46
Subject = 47
beta-carotene (solid)
Vitamin E (dashed)
months since randomization
0
1000
2000
3000
0
2
4
6
8
10
12
0
5
10
15

Dose = 45


Subject = 9
Subject = 10
Subject = 12
Subject = 13
Subject = 23
Subject = 31
Subject = 46
Subject = 47
log(beta-carotene) (solid)
Vitamin E (dashed)
months since randomization

Categorical Longitudinal Data

Example 3: Maternal Stress and Child Morbidity

- daily indicators of stress (maternal), and illness (child)
- primary outcome: illness, utilization
- covariates: employment, stress
- **Q**: association between employment, stress and morbidity?

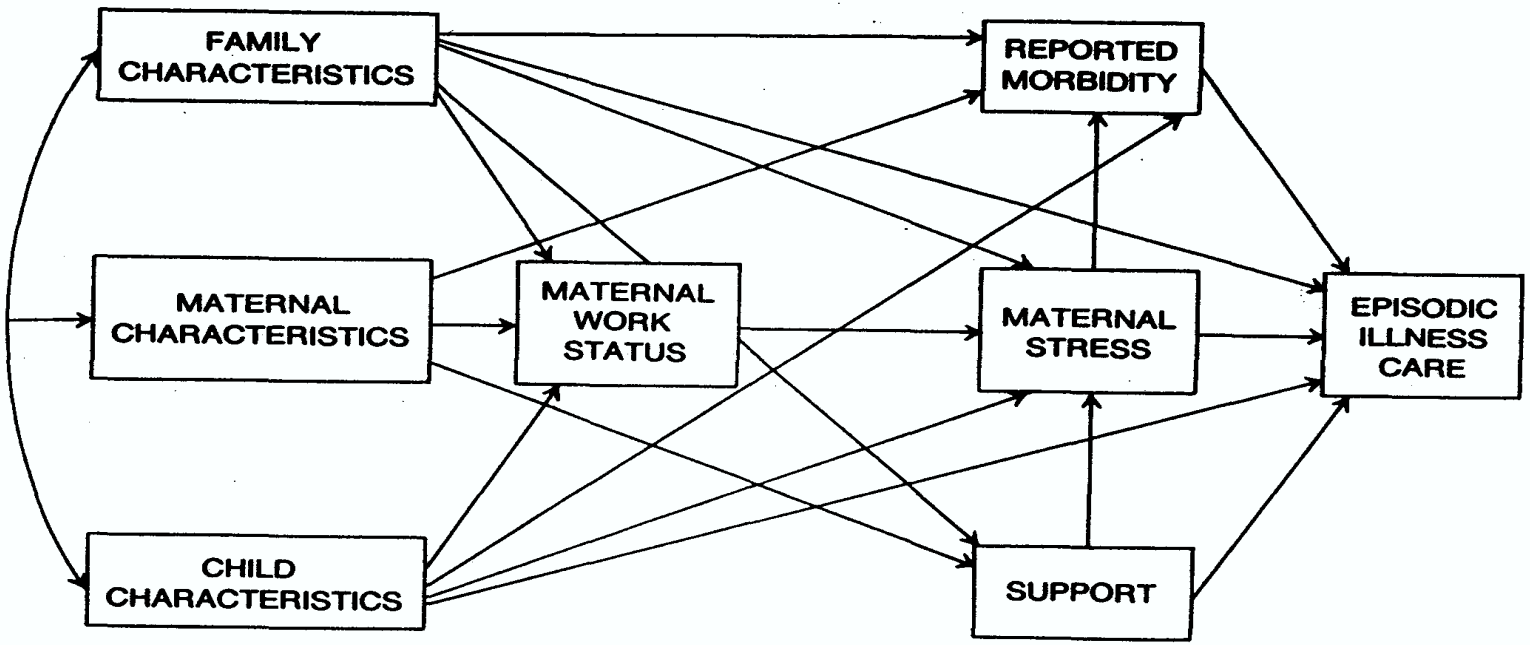
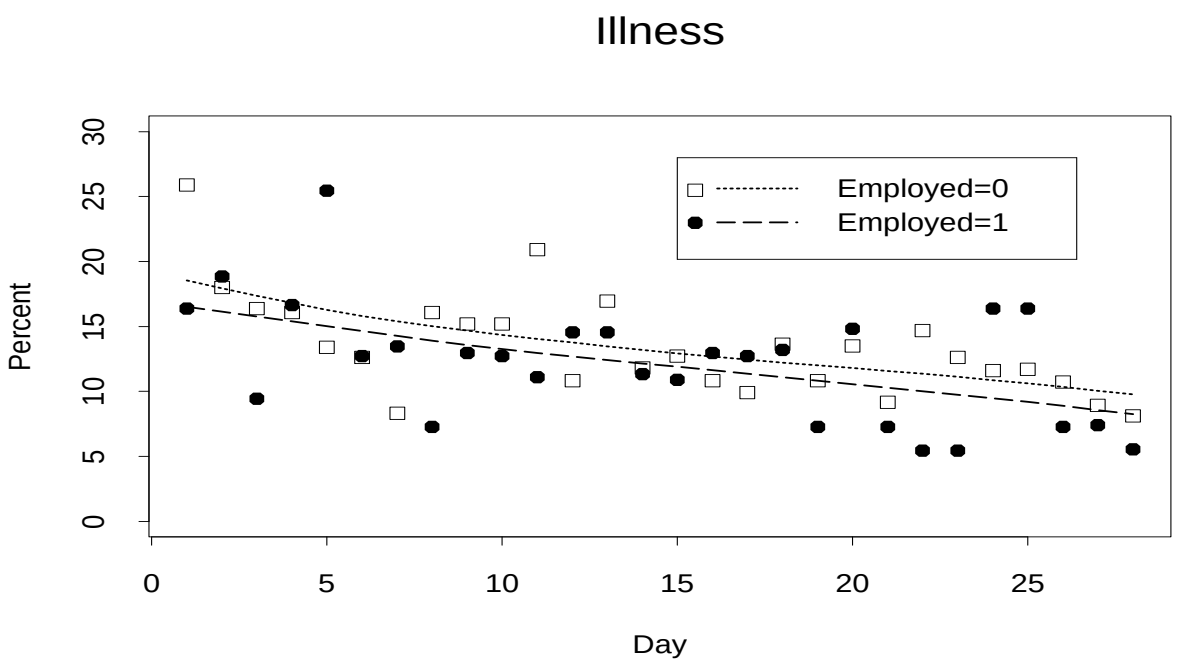
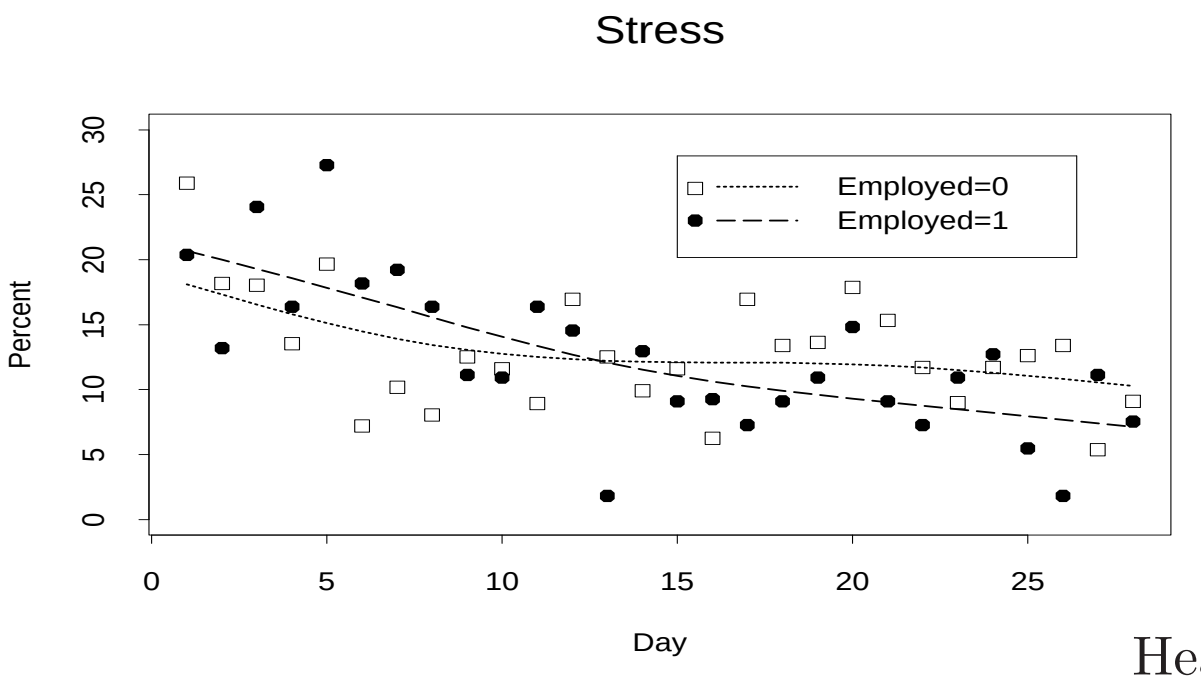
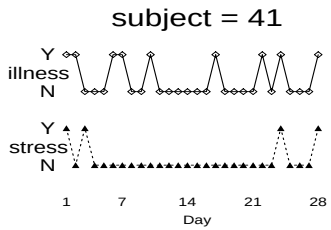
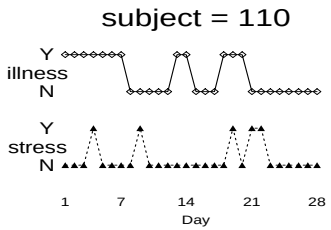


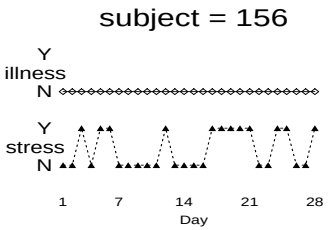
FIG. 1. Determinants of episodic illness care utilization.

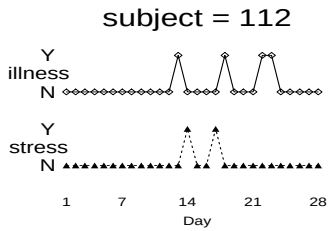

Illness
Percent
Day
Employed=0
Employed=1
0
5
10
15
20
25
30

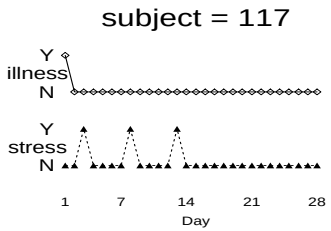

Stress
Percent
Day
Employed=0
Employed=1
0
5
10
15
20
25
30

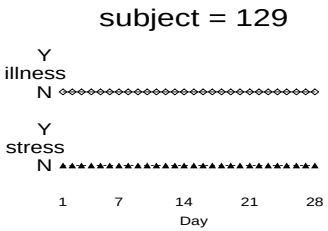

subject = 41
illness
stress
Y
N
1 7 14 21 28
Day

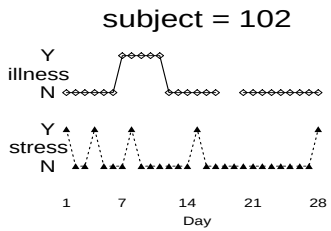

subject = 110
illness
stress
Y
N
1 7 14 21 28
Day

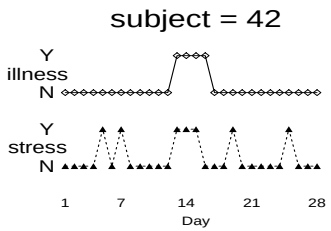

subject = 156
illness
stress
Y
N
1 7 14 21 28
Day

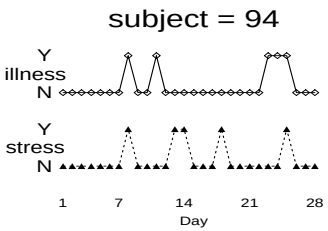

subject = 112
illness
stress
Y
N
1 7 14 21 28
Day

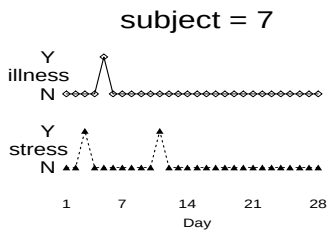

subject = 117
illness
stress
Y
N
1 7 14 21 28
Day

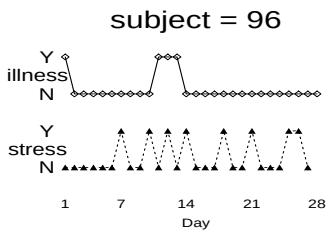

subject = 129
illness
stress
Y
N
1 7 14 21 28
Day

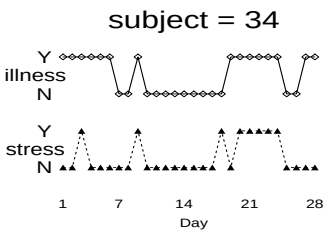

subject = 102
illness
stress
Y
N
1 7 14 21 28
Day


subject = 42
illness
stress
Y
N
1 7 14 21 28
Day


subject = 94
illness
stress
Y
N
1 7 14 21 28
Day


subject = 7
illness
stress
Y
N
1 7 14 21 28
Day

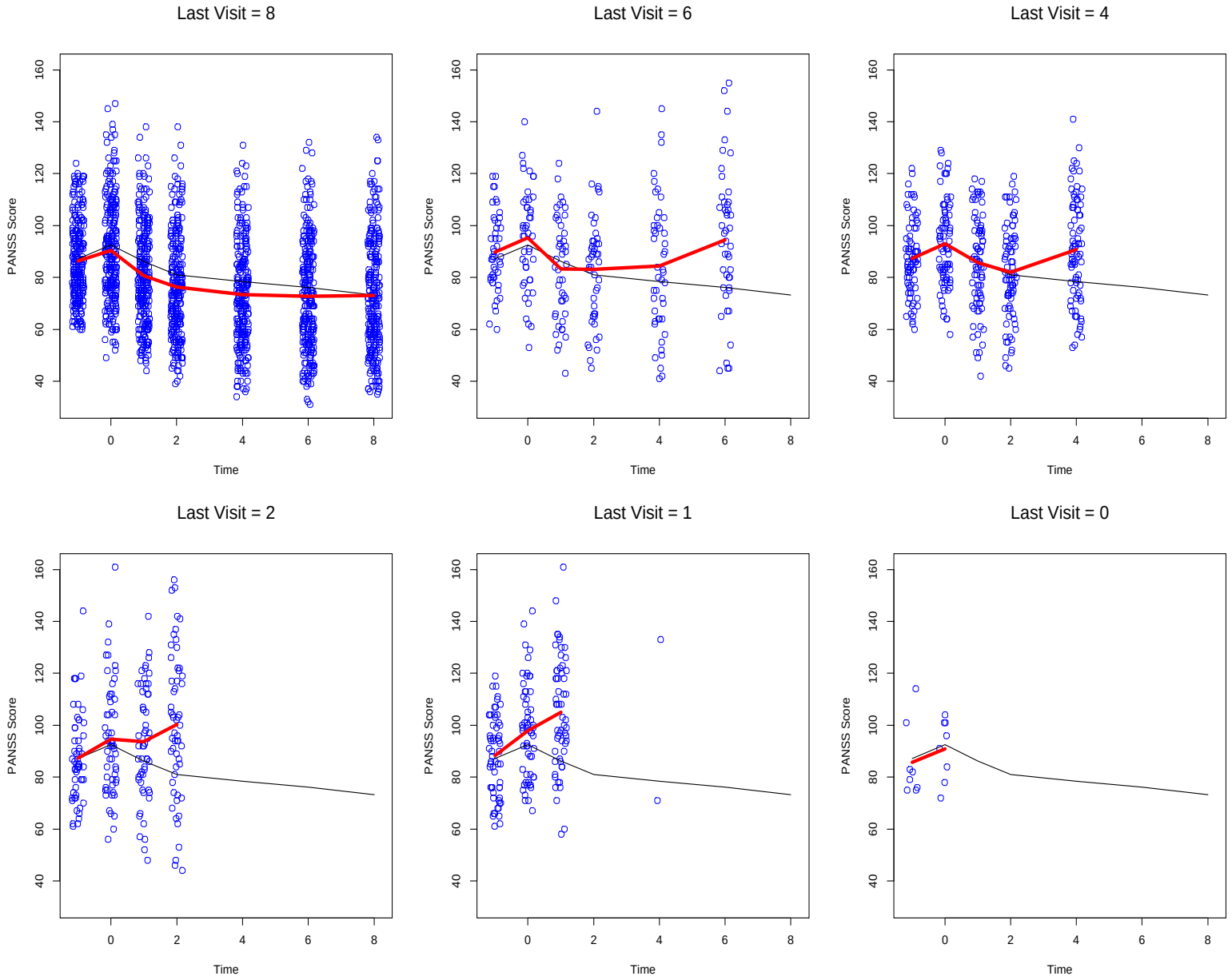

subject = 96
illness
stress
Y
N
1 7 14 21 28
Day


subject = 34
illness
stress
Y
N
1 7 14 21 28
Day

Continuous Longitudinal Data

Example 4: PANSS Data

- PANSS is a standard symptom assessment for schizophrenic patients. This study compares different doses of a new agent to a standard agent and to placebo.
- primary outcome: PANSS
- covariates: treatment, time.
- **Q**: What's the best treatment?


Last Visit = 8
Last Visit = 6
Last Visit = 4
Last Visit = 2
Last Visit = 1
Last Visit = 0
PANSS Score
Time

Longitudinal Data

- In longitudinal studies of health, we typically observe two distinct kinds of outcomes
  - ▷ **Times** of clinical or other key events
  - ▷ **Repeated values** of *markers* of the health status of participants
- In general terms, the scientific question is how explanatory variables affect times to clinical events *and* markers of *the level or change* in health status over time
- The relationship between the event times and markers can also be of interest
  - ▷ Use markers as predictors of, or surrogates for, the clinical event time

Longitudinal Studies

Benefits of longitudinal studies:

1. **Incident events are recorded**
   - Measure the new occurrence of disease.
   - Timing of disease onset can be correlated with recent changes in patient exposure and/or with chronic exposure.
2. **Prospective ascertainment of exposure**
   - Participants can have their exposure status recorded at multiple follow-up visits. This can alleviate recall bias.
   - Temporal order of exposures and outcomes is observed.

3. **Measurement of individual change in outcomes**
   - A key strength of a longitudinal study is the ability to measure change in outcomes and/or exposure at the individual level.
   - Longitudinal studies provide the opportunity to observe individual patterns of change.
4. **Separation of time effects: Cohort, Period, Age**
   - When studying change over time there are many time scales to consider.
     - **cohort** scale is the time of birth such as 1945 or 1963.
     - **period** is the current time such as 2004.
     - **age** is (period - cohort).
   - A longitudinal study with times $t_1, t_2, \ldots t_n$ can characterize multiple time scales such as age and cohort effects using covariates derived from the calendar time and birth year: age of subject i at time t_j is $\text{age}_{ij} = (t_j - \text{birth}_i)$; and cohort is $\text{cohort}_{ij} = \text{birth}_i$.

5. Control for cohort effects

- In a cross-sectional study the comparison of subgroups of different ages combines the effects of aging and the effects of different cohorts. That is, comparison of outcomes measured in 2003 among 58 year old subjects and among 40 year old subjects reflects both the fact that the groups differ by 18 years (aging) and the fact that the subjects were born in different eras.
- In a longitudinal study the cohort under study is fixed and thus changes in time are not confounded by cohort differences.

An nice overview of LDA opportunities in respiratory epidemiology is presented in Weiss and Ware (1996). Lebowitz (1996) discusses age, period, and cohort effects.

Longitudinal Studies

The benefits of a longitudinal design are not without cost. There are several challenges posed:

Challenges of longitudinal studies:

1. **Participant follow-up**

   Risk of bias due to incomplete follow-up, or "drop-out" of study participants. If subjects that are followed to the planned end of study differ from subjects who discontinue follow-up then a naive analysis may provide summaries that are not representative of the original target population.

2. Analysis of correlated data

- Statistical analysis of longitudinal data requires methods that can properly account for the intra-subject correlation of response measurements.
- If such correlation is ignored then inferences such as statistical tests or confidence intervals can be grossly invalid.

3. **Time-varying covariates**

- Although longitudinal designs offer the opportunity to associate changes in exposure with changes in the outcome of interest, the direction of causality can be complicated by "feedback" between the outcome and the exposure.
- Example = MSCM with stresss and illness.
- Although scientific interest generally lies in the effect of exposure on health, reciprocal influence between exposure and outcome poses analytical difficulty when trying to separate the effect of exposure on health from the effect of health on exposure.
- How to choose exposure "lag"?
  - e.g. Is it the air pollution today, yesterday, or last week that is the important predictor of morbidity today?

Longitudinal Studies

The Scientific Opportunity

- Observe individual **changes** over time.
- Characterize the time-course of disease.

Outcome Measures

- A single outcome at a fixed follow-up time.
- The time until an event occurs.

$\star\star\star$ Repeated measures taken over time.

Motivation

Cystic Fibrosis and Pulmonary Function

- Several specific aspects are of interest:

  1. What is the rate of decline in FEV1?
  2. Is the time course different for males and females?
  3. Is the time course different for F508 homozygous subjects ?

- **Reference**: Davis P.B. (1997) *Journal of Pediatrics*

Data

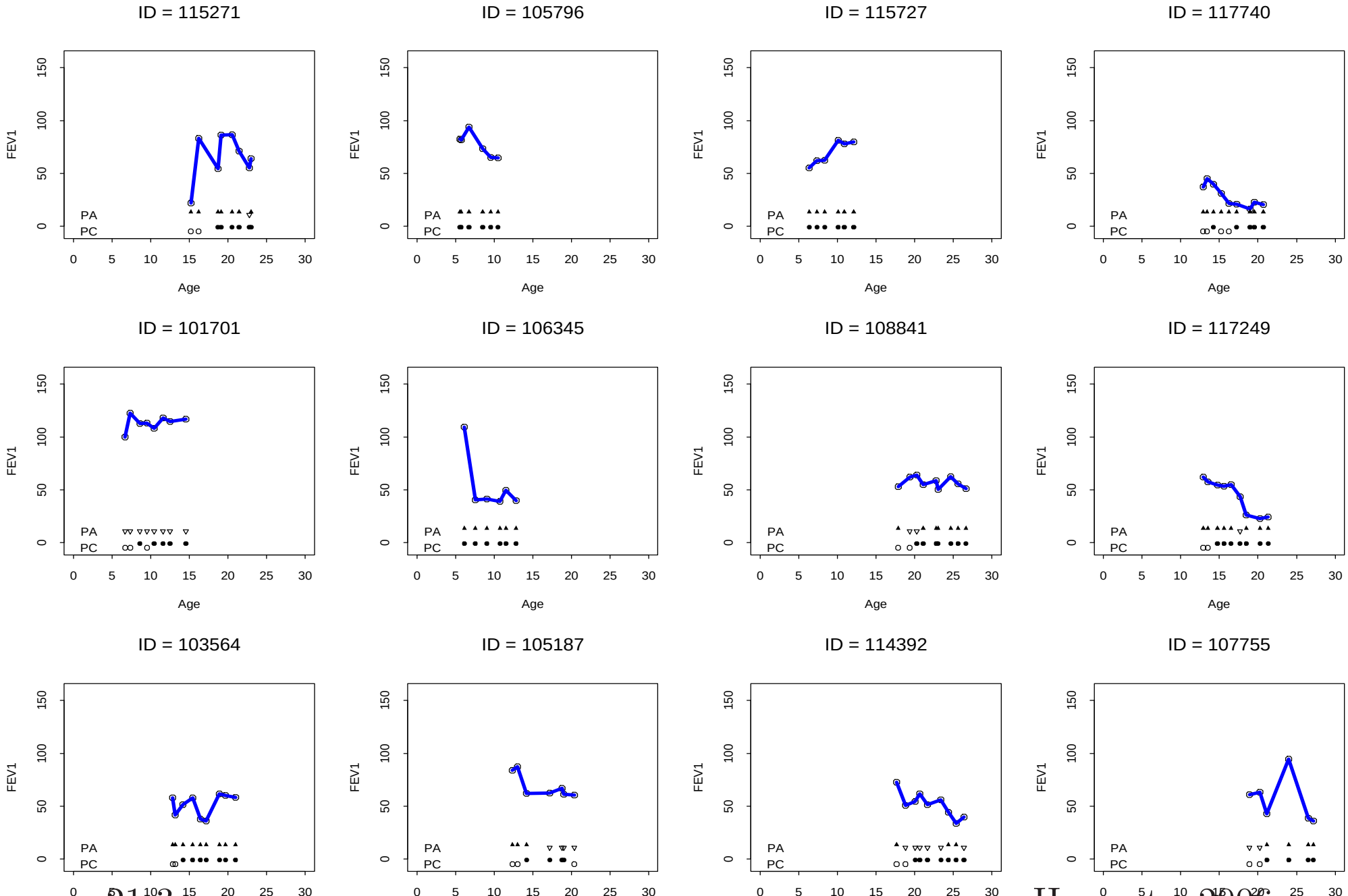
```
  ID       = patient id
  FEV1     = percent-predicted forced expiratory volume in 1 second
  AGE      = age (years)
  GENDER   = sex (1=male, 2=female)
  PSEUDOA  = infection with Pseudomonas Aeruginosa (0=no, 3=yes)
  F508     = genotype (1=homozygous, 2=heterozygous, 3=none)
  PANCREAT = pancreatic enzyme supplmentation (0,1=no, 2=yes)

100073 113.8   8.452 2 3 1 2
100073  98.18  8.783 2 3 1 2
100073  98.73  9.785 2 3 1 2
100073 101.79 10.538 2 3 1 2
100073  98.04 12.329 2 3 1 2
100073  94.32 13.306 2 3 1 2
100073  95.48 14.418 2 3 1 2
100111  96.85 12.515 2 0 3 1
100111 101.05 13.103 2 0 3 2
100111 100.33 15.105 2 0 3 2
100111  90.92 16.838 2 0 3 2
100111 109.78 17.582 2 0 3 2
100111 107.76 18.847 2 0 3 1
```

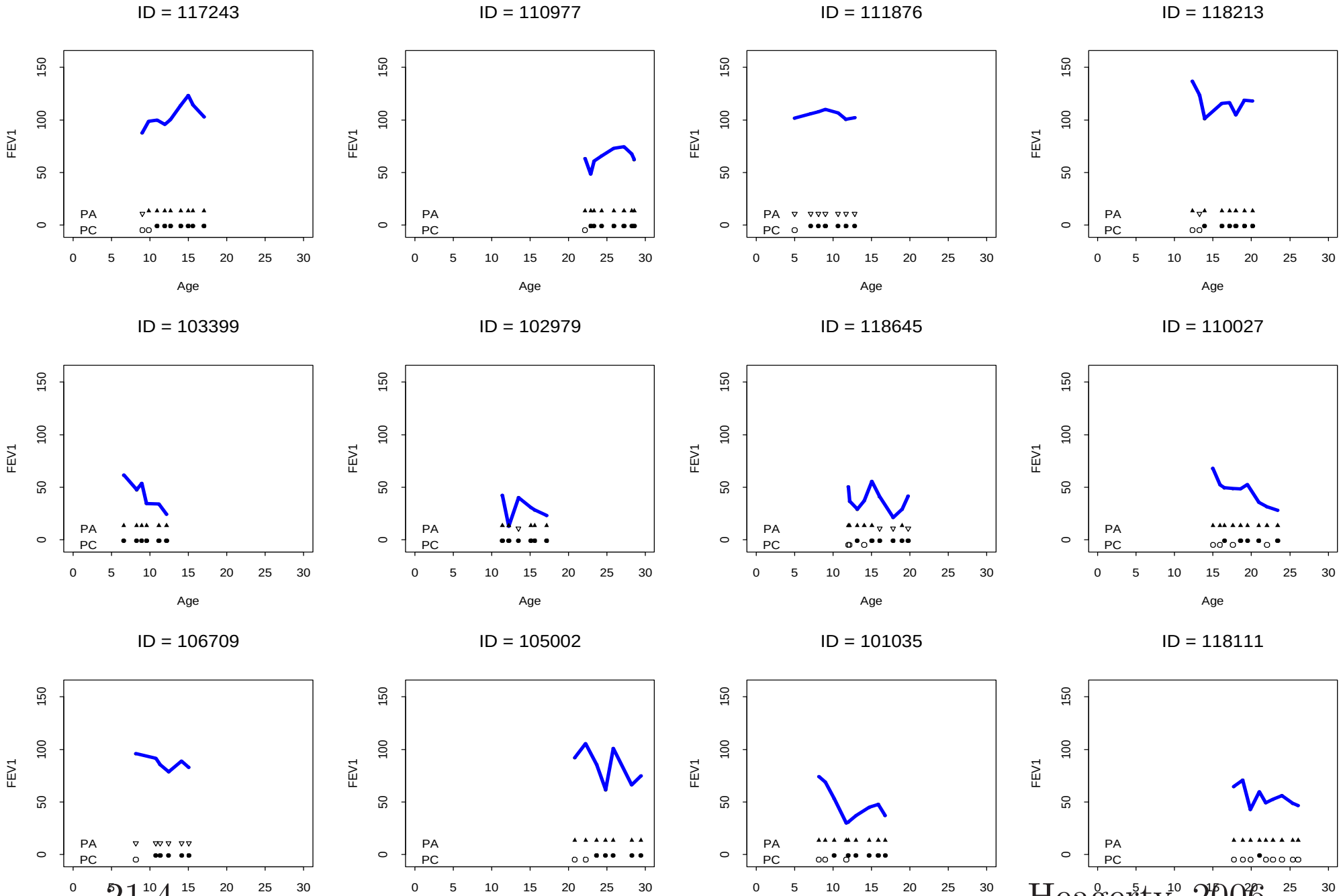
EDA: Numerical Summaries

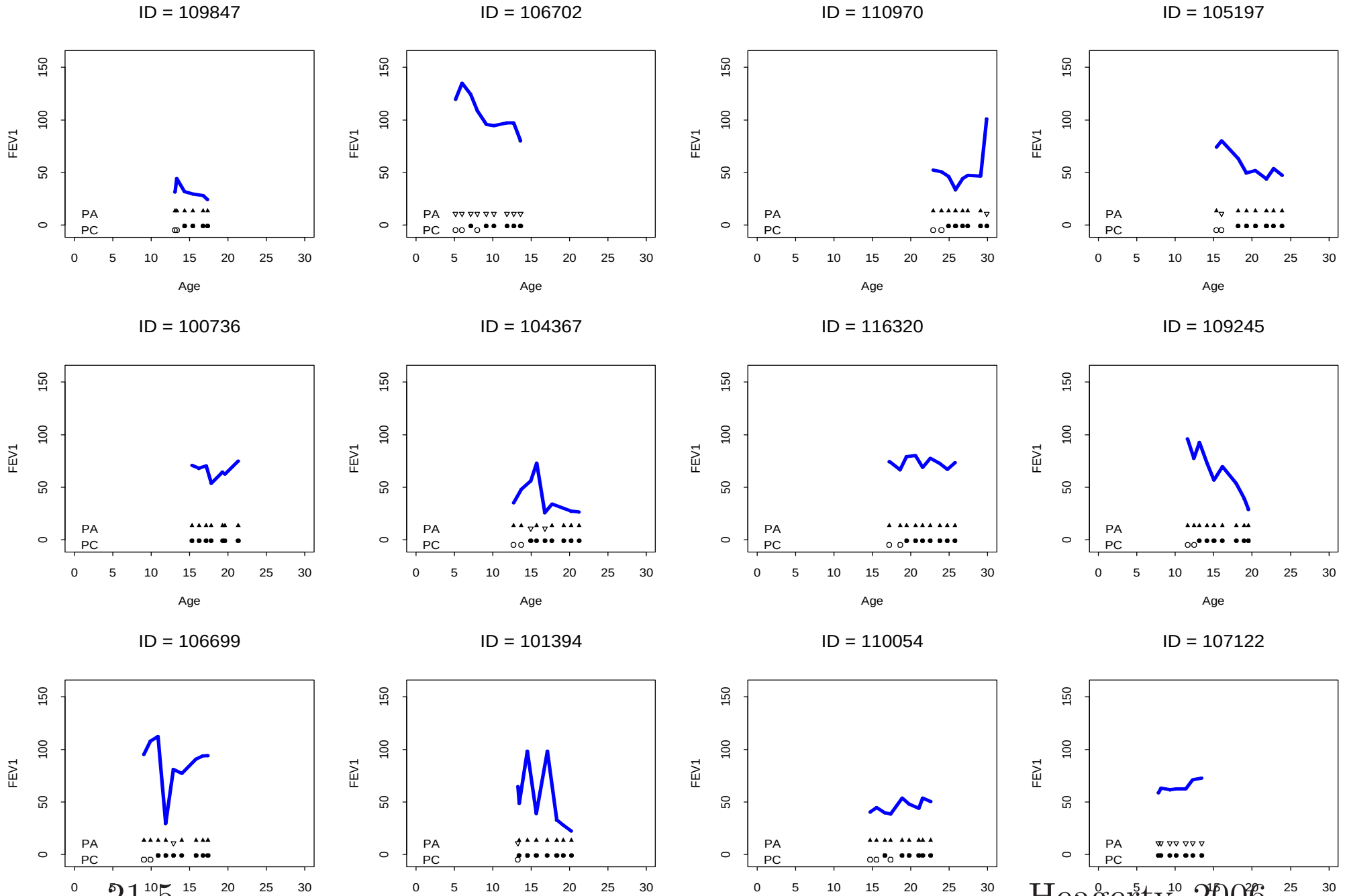
```
Total number of subjects = 200


Number of observations (number of subjects with ni):
  6  7  8  9
 49 52 36 63


Distribution of males / females
 male female
  102     98


Number of mutations of f508
  0  1  2
 23 87 90
```

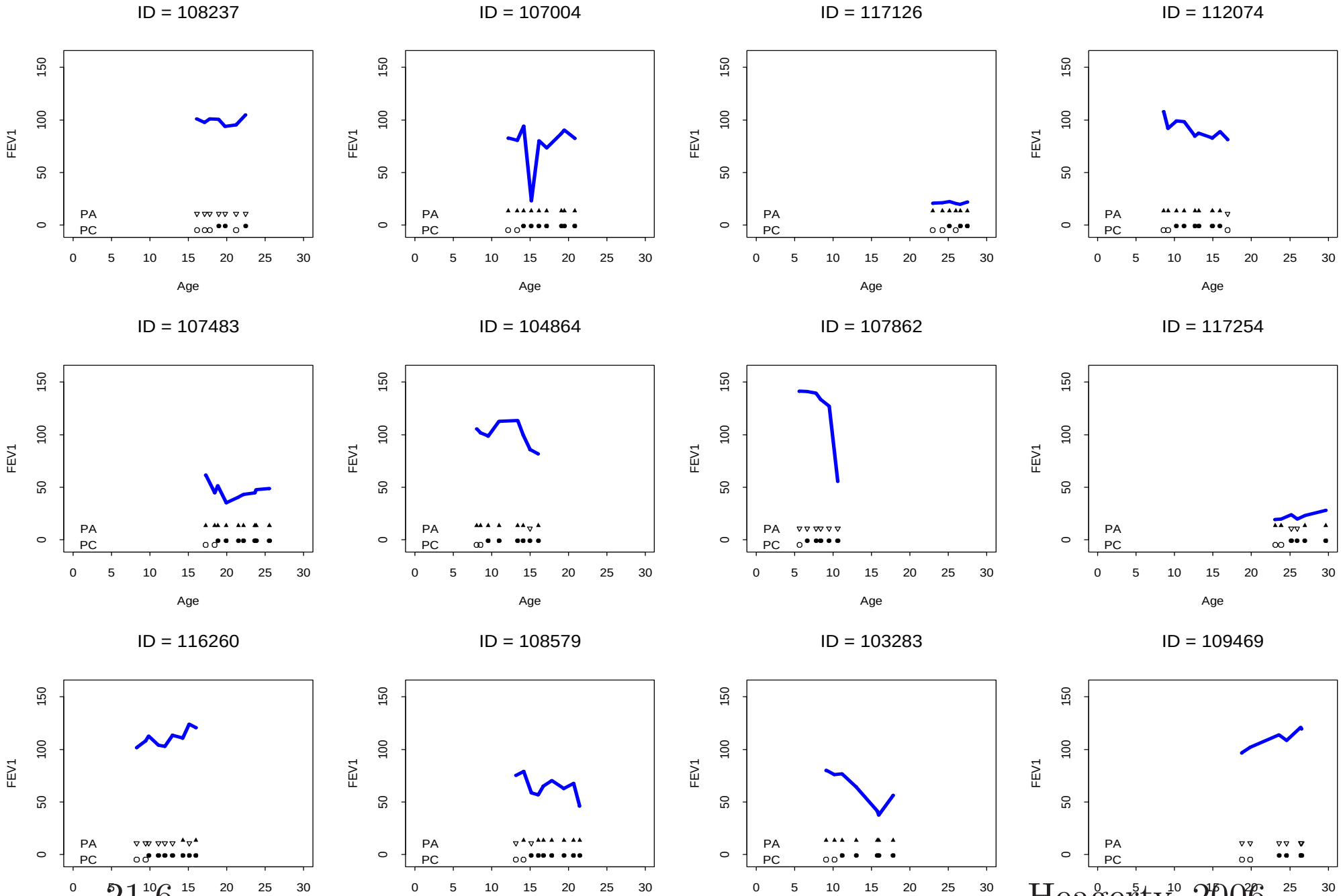
EDA: Numerical Summaries

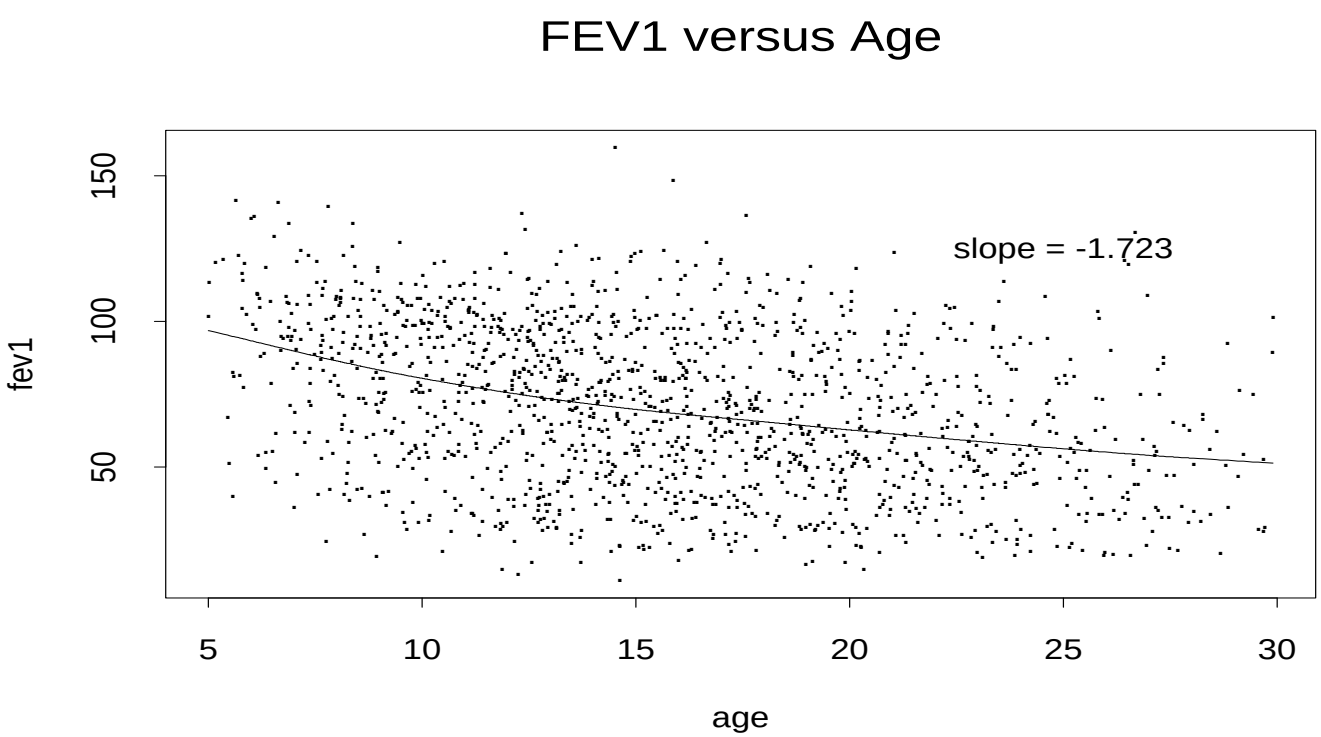
```
Age at entry

N = 200   Median = 11.9655
Quartiles = 7.758, 15.3235

Decimal point is at the colon

    5 : 002355666788889
    6 : 0111222334555567789999
    7 : 0001234446778889
    8 : 011223345566899
    9 : 00011244788
   10 : 0111113349
   11 : 2223446678
   12 : 00111222334455577888888999
   13 : 01234455
   14 : 111245555779
   15 : 001223357
   16 : 0012347899
   17 : 1223567779
   18 : 4899
   19 : 4
   20 : 0123778
   21 : 15577
   22 : 2459
   23 : 001128
```

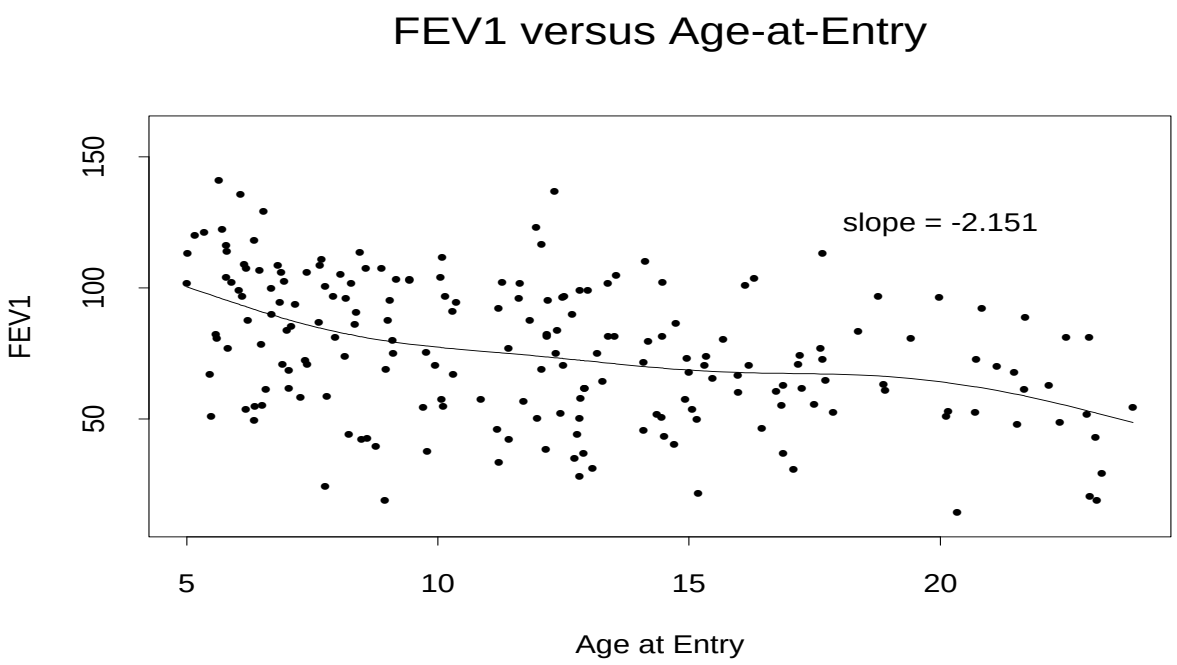

ID = 115271
ID = 105796
ID = 115727
ID = 117740
ID = 101701
ID = 106345
ID = 108841
ID = 117249
ID = 103564
ID = 105187
ID = 114392
ID = 107755
FEV1
Age
PA
PC

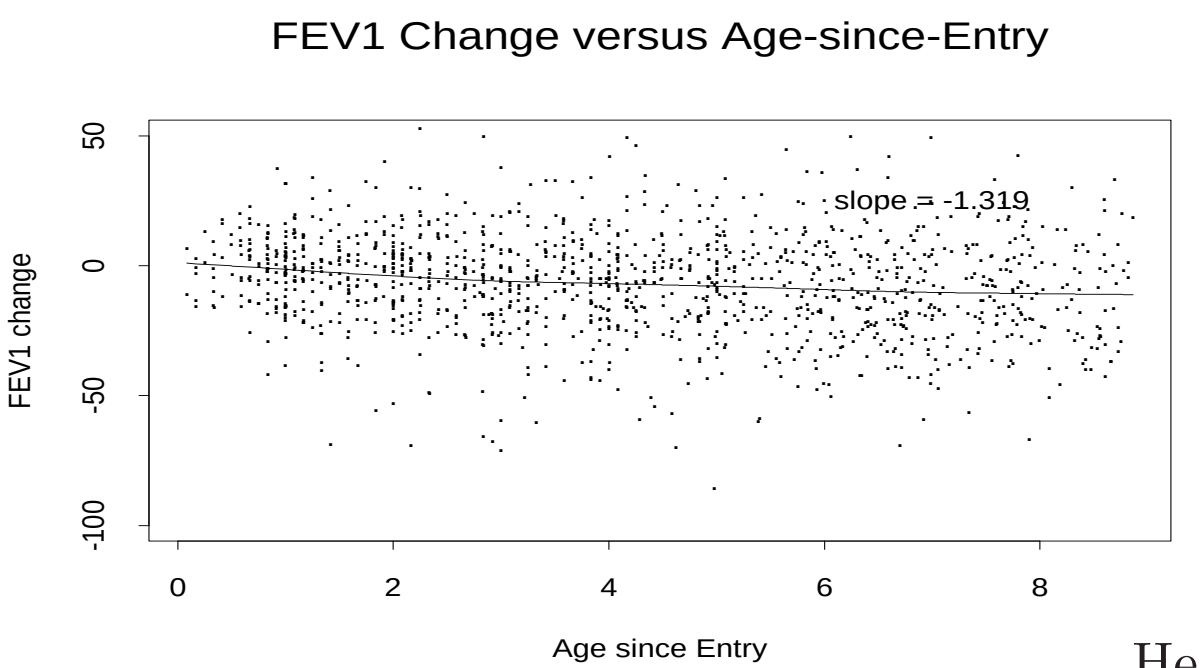

ID = 117243
ID = 110977
ID = 111876
ID = 118213
ID = 103399
ID = 102979
ID = 118645
ID = 110027
ID = 106709
ID = 105002
ID = 101035
ID = 118111
FEV1
Age
PA
PC


ID = 109847
ID = 106702
ID = 110970
ID = 105197
ID = 100736
ID = 104367
ID = 116320
ID = 109245
ID = 106699
ID = 101394
ID = 110054
ID = 107122
FEV1
Age
PA
PC


ID = 108237
ID = 107004
ID = 117126
ID = 112074
ID = 107483
ID = 104864
ID = 107862
ID = 117254
ID = 116260
ID = 108579
ID = 103283
ID = 109469
FEV1
Age
PA
PC


FEV1 versus Age
slope = -1.723
fev1
150
100
50
5
10
15
20
25
30
age


FEV1 versus Age-at-Entry
slope = -2.151
FEV1
150
100
50
5
10
15
20
Age at Entry


FEV1 Change versus Age-since-Entry
slope = -1.319
FEV1 change
50
0
-50
-100
0
2
4
6
8
Age since Entry

Distinguishing Cross-sectional and Longitudinal Associations

- Cross-sectional data

$$Y_{i1} = \beta_C x_{i1} + \epsilon_{i1}, \quad i = 1, \ldots, m \tag{1}$$

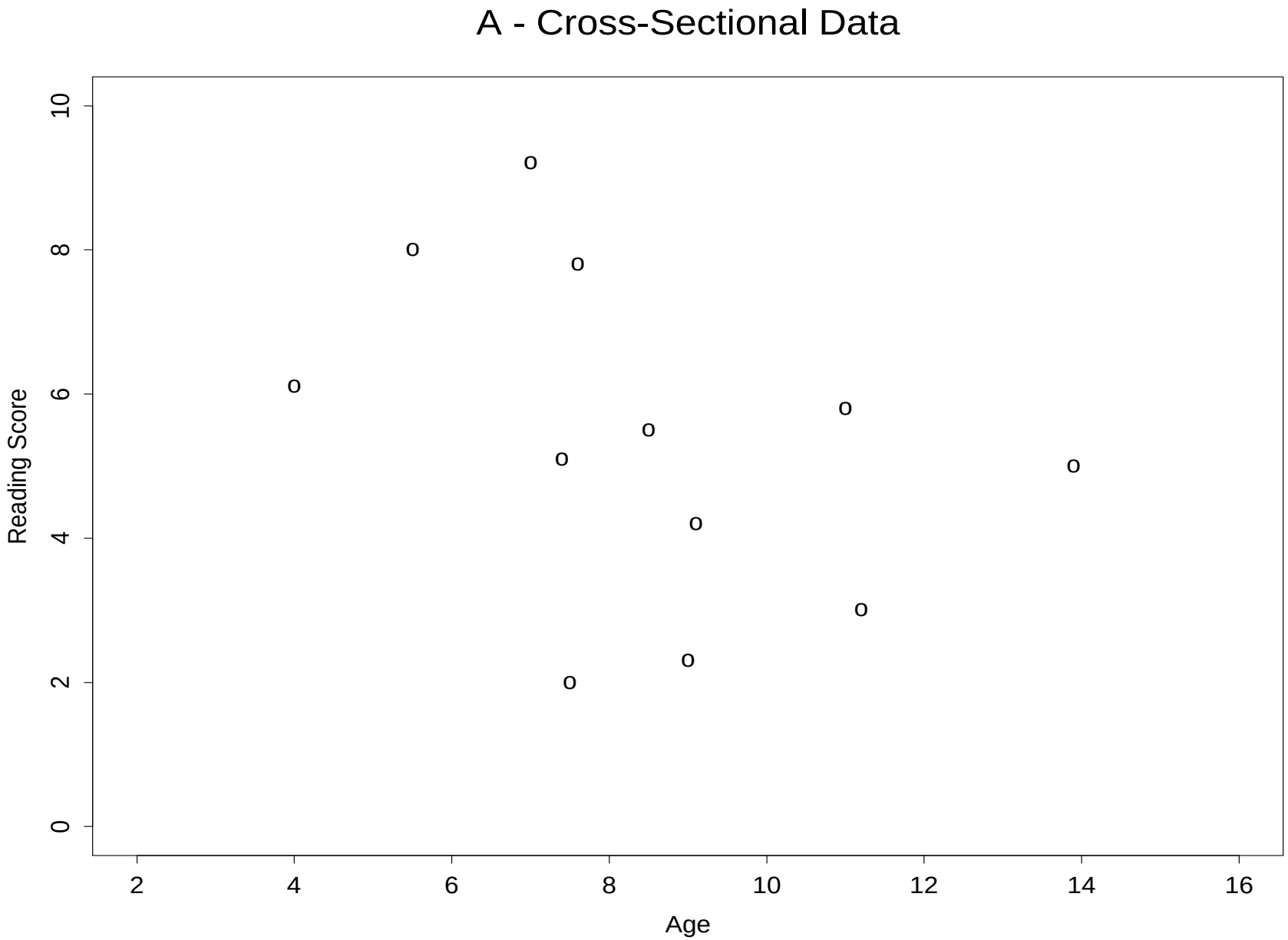
- β_C represents the difference in average Y across two sub-populations which differ by one unit in x.
- Longitudinal data

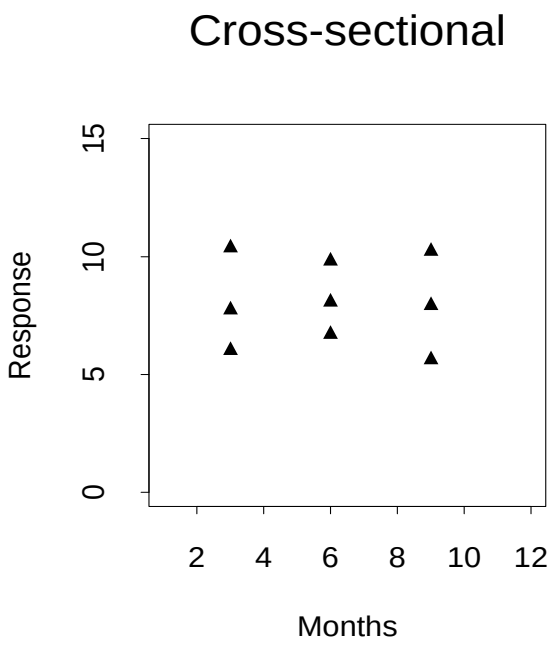
$$Y_{ij} = \beta_C x_{i1} + \beta_L (x_{ij} - x_{i1}) + \epsilon_{ij}, \quad \begin{matrix} j = 1, \ldots, n_i \\ i = 1, \ldots, m \end{matrix} \tag{2}$$

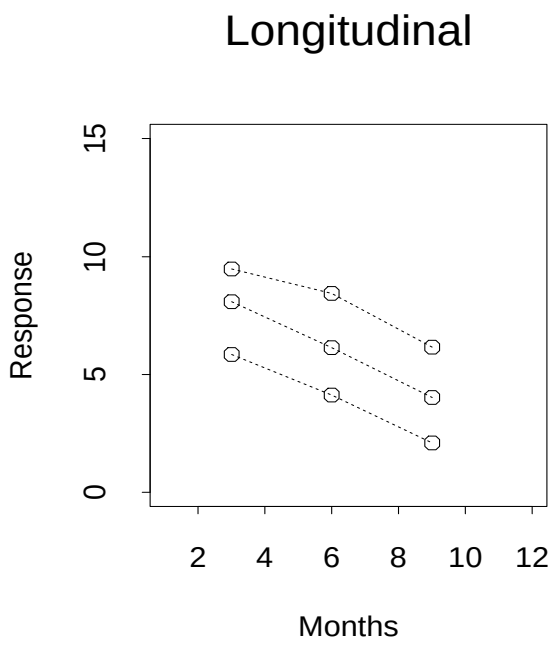
- When $j = 1$, the two equations are the same; β_C has the same cross-sectional interpretation
- Subtract equations above to obtain

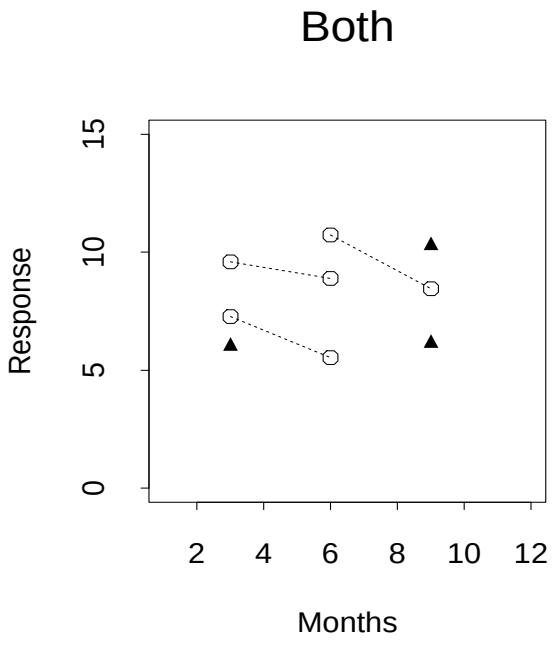
$$(Y_{ij} - Y_{i1}) = \beta_L(x_{ij} - x_{i1}) + (\epsilon_{ij} - \epsilon_{i1}).$$

- β_L represents the expected **change** in Y per unit **change** in x

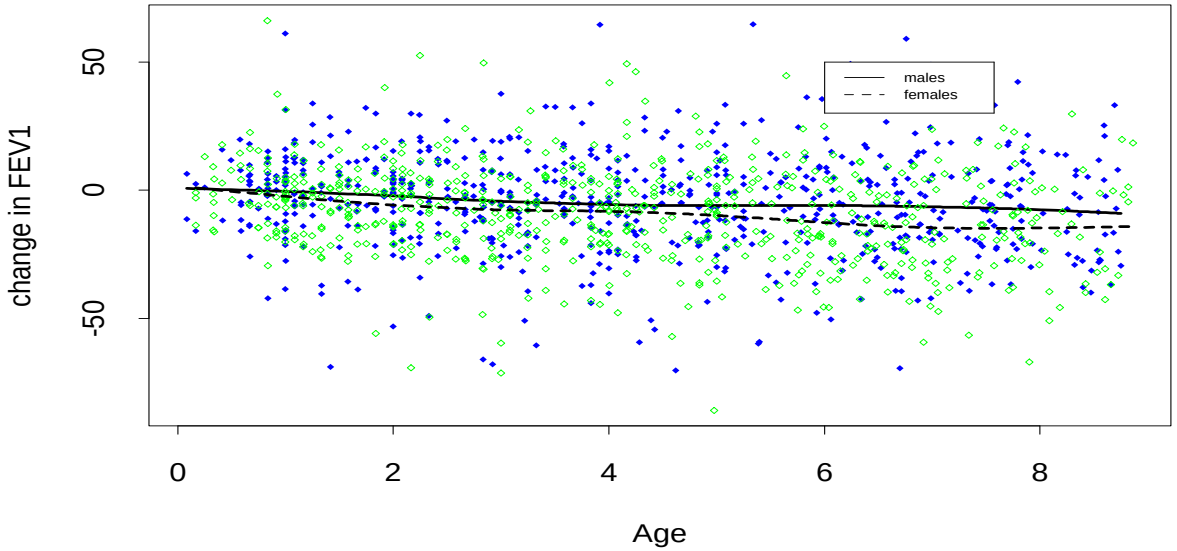

A - Cross-Sectional Data
Reading Score
Age
0
2
4
6
8
10
2
4
6
8
10
12
14
16


Cross-sectional
Response
0
5
10
15
2
4
6
8
10
12
Months

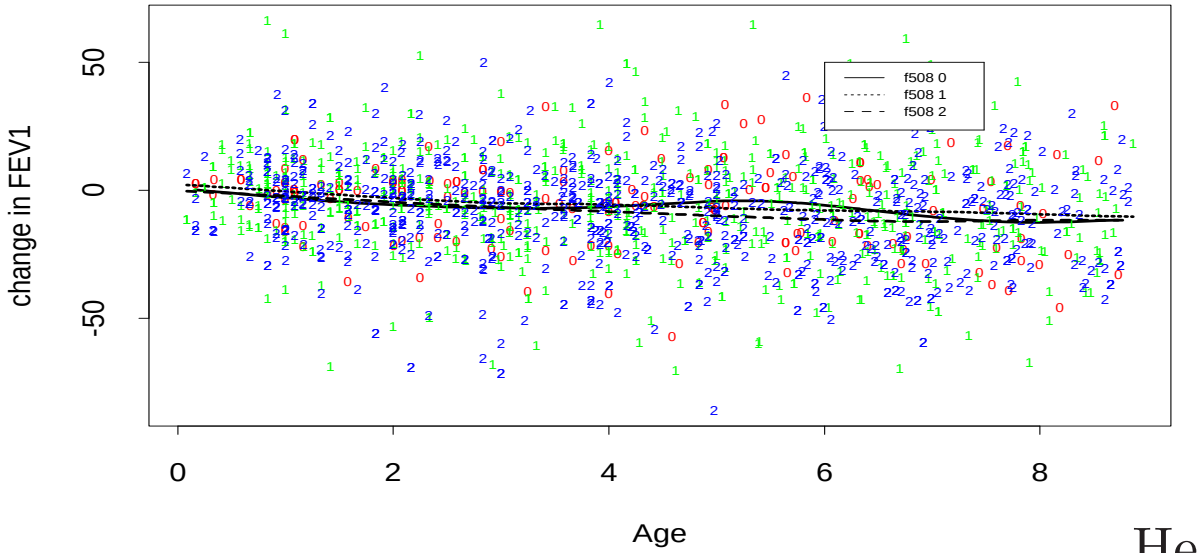

Longitudinal
Response
0
5
10
15
2
4
6
8
10
12
Months


Both
Response
0
5
10
15
2
4
6
8
10
12
Months

FEV1 by Male/Female



FEV1 by f508



EDA Summary

Observations

- Systematic trends: time, gender, F508.
- Random variation: individual, observation.

Questions

- Two time scales?
- Estimation / testing for rates of decline?
- Other?

Longitudinal Data Analysis

INTRODUCTION

to

CORRELATION and WEIGHTING

Longitudinal Data

"The basic statistical problem is that variables from a given individual are correlated over time." $(generic)$

Q: So what?

- (-) ignoring dependence can lead to invalid inference.
- (-) often limited information regarding dependence.
- (+) can observe **change** for individuals over time.
- (+) variety of statistical approaches that are available.

Longitudinal Data

"... need to account for the dependence." (*generic*)

Q: How?

1. **Choice of Model**
2. **Choice of Estimator**
3. **Choice of Summaries**

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 $i = (n + 1) : 2n$

$$\hat{\mu}_0 = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2} Y_{ij}$$

$$\hat{\mu}_1 = \frac{1}{2n} \sum_{i=n+1}^{2n} \sum_{j=1}^{2} Y_{ij}$$

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

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

(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 $i = (n + 1) : 2n$

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

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

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

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 dependence 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:

Data : $(1, 10)$ $(2, 100)$

$$\hat{p}_1 = (2+1)/(110) = 0.030$$

$$\hat{p}_2 = \frac{1}{2}\{1/10 + 2/100\} = 0.051$$

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 $1/\text{variance of } Y_i/N_i$ (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

- Longitudinal (dependent) data are common (and interesting!).
- Inference must account for the dependence.
- Consideration as to the choice of weighting will depend on the variance/covariance of the response variables.

Summary

Methodological Issues in the Study of Cognitive Decline

Morris et al. AJE (1999)

"In modeling associations between risk factors and change in cognitive function, the usual analytic issues are complicated by the need to consider time in the analysis. The potential for both fruitful investigation and misinterpretation of the data is increased. Analyses that make the most effective use of the data usually require the most advanced methods of longitudinal analysis. "

Summary

Methodological Issues in the Study of Cognitive Decline

Morris et al. AJE (1999)

"Use of the longitudinal designs and advanced statistical models described here is well worth the extra effort required. Meeting the special challenges of measuring change in cognitive function should improve our ability to identify risk factors for cognitive decline and other diseases related to aging. Many of these issues apply to any study of disease processes entailing change with age."

Some References: Books

Diggle PJ, Heagerty PJ, Liang K-Y, Zeger SL (2002) *Analysis of Longitudinal Data, Second Edition*, Oxford University Press.

Fitzmaurice GM, Laird NM, Ware JM (2004) *Applied Longitudinal Analysis*, Wiley.

Verbeke G, Molenberghs G (2000) *Linear Mixed Models for Longitudinal Data*, Springer.

Singer JD, Willett JB (2003) *Applied Longitudinal Data Analysis*, Oxford University Press.

Longitudinal Data Analysis

INTRODUCTION

to

REGRESSION APPROACHES

Linear Mixed Model

- **Regression model:**
  mean response as a function of covariates.
  "systematic variation"

- **Random effects:**
  variation from subject-to-subject in trajectory.
  "random between-subject variation"

- **Within-subject variation:**
  variation of individual observations over time
  "random within-subject variation"

Scientific Questions as Regression

$\star$ Questions concerning the rate of decline refer to the time slope for FEV1:

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

Time Scales

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

CF Regression Model

Model:

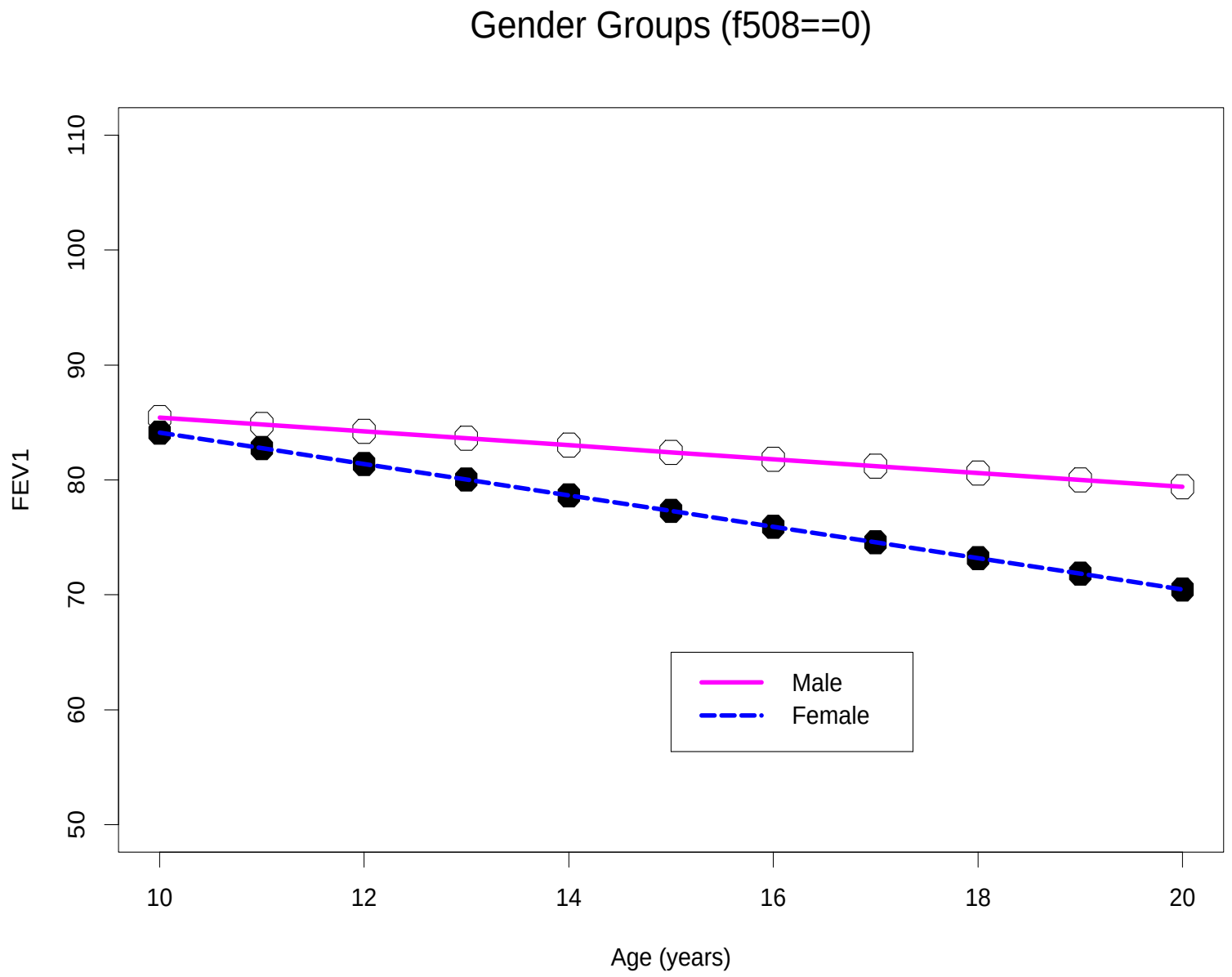
$$
\begin{aligned}
E[\texttt{FEV} \mid \boldsymbol{X}_i] &= \beta_0 \\
&\quad + \beta_1 \cdot \texttt{age0} + \beta_2 \cdot \texttt{ageL} \\
&\quad + \beta_3 \cdot \texttt{female} \\
&\quad + \beta_4 \cdot \texttt{f508} = 1 + \beta_5 \cdot \texttt{f508} = 2 \\
&\quad + \beta_6 \cdot \texttt{female} \cdot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

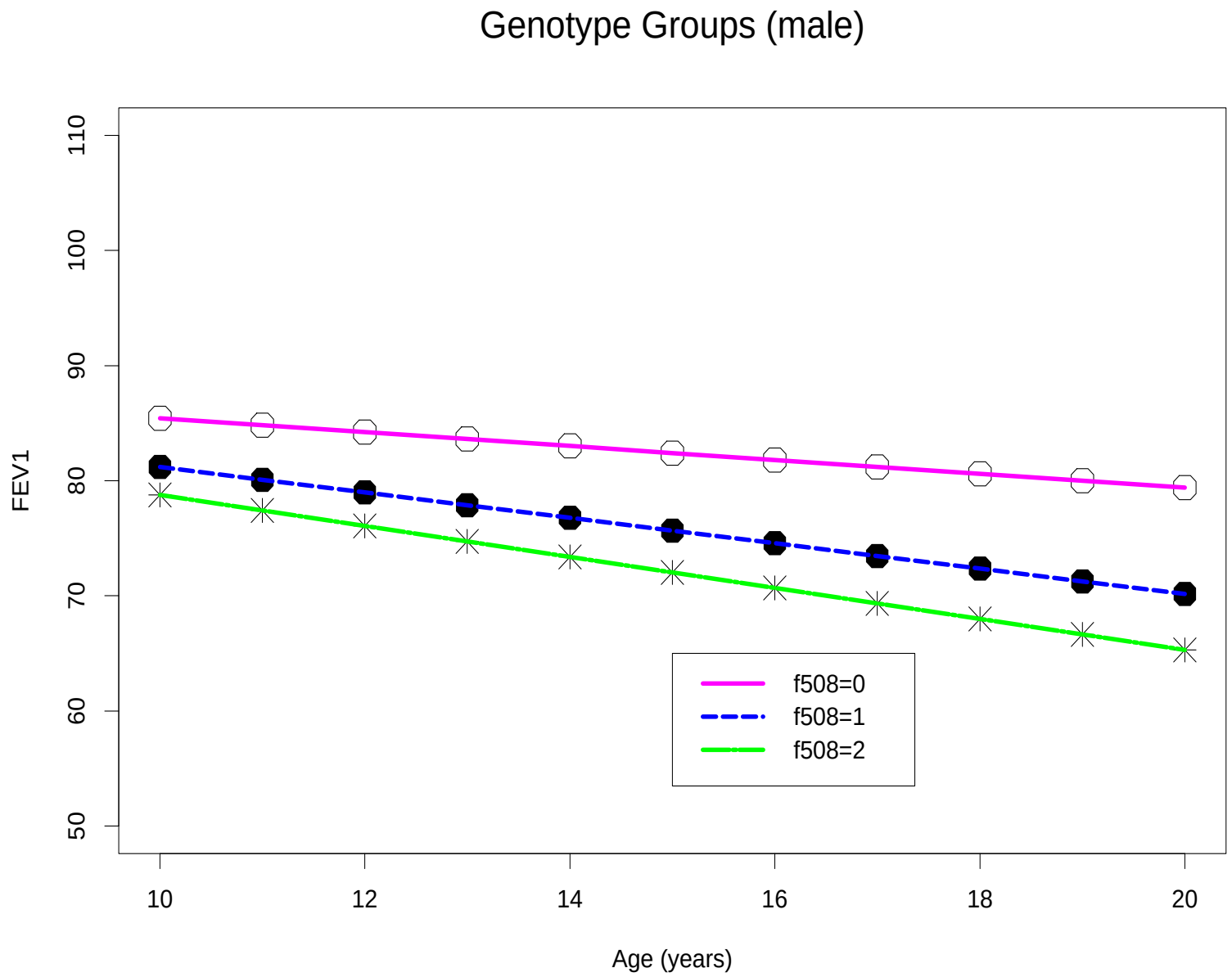
Intercept

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

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

Define

$$Y_{ij} = \text{FEV1 for subject } i \text{ at time } t_{ij}$$

$$\boldsymbol{X}_i = (\boldsymbol{X}_{ij}, \ldots, \boldsymbol{X}_{in_i})$$

$$\boldsymbol{X}_{ij} = (X_{ij,1}, X_{ij,2}, \ldots, X_{ij,p})$$

age0, ageL, gender, genotype

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.
- 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 not 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 is predictive of the LDL cholesterol of patient ID=212 at age 60.

Dependence and Correlation

- 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

- 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

- 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 pages 28 and 29 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...

$$\text{cov}(Y_i) = \begin{bmatrix} \text{var}(Y_{i1}) & \text{cov}(Y_{i1}, Y_{i2}) & \dots & \text{cov}(Y_{i1}, Y_{in_i}) \\ \text{cov}(Y_{i2}, Y_{i1}) & \text{var}(Y_{i2}) & \dots & \text{cov}(Y_{i2}, Y_{in_i}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(Y_{in_i}, Y_{i1}) & \text{cov}(Y_{in_i}, Y_{i2}) & \dots & \text{var}(Y_{in_i}) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \dots & \sigma_1\sigma_{n_i}\rho_{1n_i} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2^2 & \dots & \sigma_2\sigma_{n_i}\rho_{2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n_i}\sigma_1\rho_{n_i1} & \sigma_{n_i}\sigma_2\rho_{n_i2} & \dots & \sigma_{n_i}^2 \end{bmatrix}$$

Mean and Covariance Models for FEV1

Models:

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

$$\text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{X}_i) \quad = \quad \boldsymbol{\Sigma}_i = \underbrace{\boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T}_{\texttt{between-subjects}} + \underbrace{\boldsymbol{R}_i}_{\texttt{within-subjects}}$$

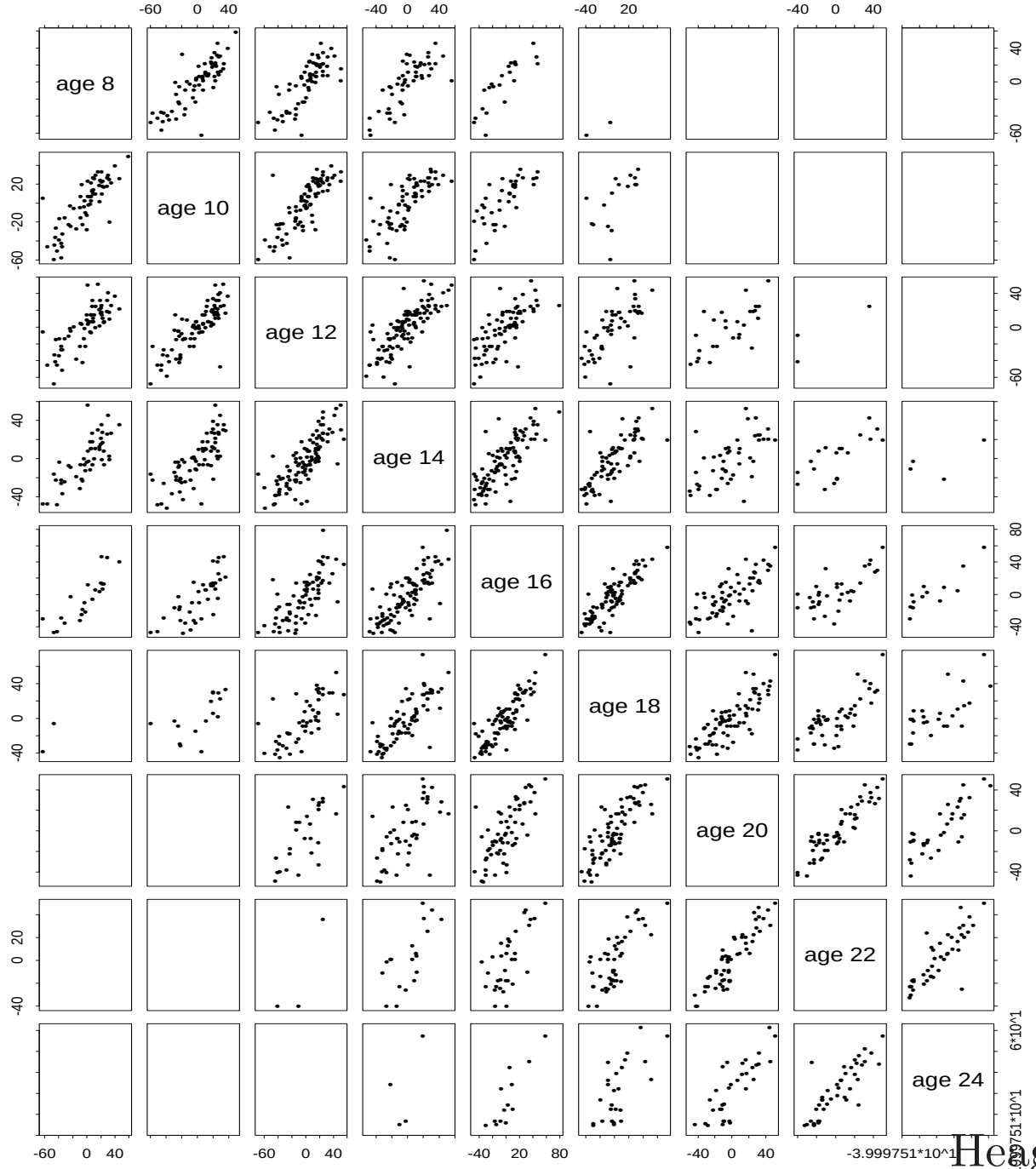
Q: What are appropriate covariance models for the FEV1 data?

"Wide Data"

Data array for the residuals (first 10 rows of ‘‘rmat’’)

	age8	age10	age12	age14	age16	age18	age20	age22	age24
[1,]	34.64	21.90	25.66	26.75	NA	NA	NA	NA	NA
[2,]	NA	NA	16.48	21.68	17.94	38.07	NA	NA	NA
[3,]	NA	NA	NA	-34.29	-34.06	-32.15	-49.71	NA	NA
[4,]	NA	NA	NA	NA	NA	NA	NA	-14.01	-6.1
[5,]	NA	NA	NA	41.59	-11.18	11.36	18.68	NA	NA
[6,]	20.55	22.71	22.78	21.55	14.61	NA	NA	NA	NA
[7,]	NA	NA	NA	NA	3.28	-8.39	3.38	19.04	NA
[8,]	NA	NA	NA	NA	-8.56	-11.92	-19.37	-24.16	NA
[9,]	39.82	39.16	36.79	NA	NA	NA	NA	NA	NA
[10,]	NA	NA	19.76	24.59	26.90	29.11	27.28	NA	NA

.
.
.


age 8
age 10
age 12
age 14
age 16
age 18
age 20
age 22
age 24

Empirical Covariance Matrix:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	714.16	575.86	515.58	470.58	648.56	1346.10	NA	NA	NA
[2,]	0.00	633.62	523.40	444.44	467.09	383.60	NA	NA	NA
[3,]	0.00	0.00	681.02	504.08	520.35	514.70	492.00	973.33	NA
[4,]	0.00	0.00	0.00	579.61	523.86	493.50	404.72	395.84	315.28
[5,]	0.00	0.00	0.00	0.00	663.82	527.98	440.31	374.98	488.25
[6,]	0.00	0.00	0.00	0.00	0.00	597.54	501.07	406.39	462.75
[7,]	0.00	0.00	0.00	0.00	0.00	0.00	605.90	487.05	511.91
[8,]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	512.90	437.70
[9,]	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	618.17

Empirical Correlation Matrix:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	1	0.86	0.74	0.73	0.94	2.06	NA	NA	NA
[2,]	0	1.00	0.80	0.73	0.72	0.62	NA	NA	NA
[3,]	0	0.00	1.00	0.80	0.77	0.81	0.77	1.65	NA
[4,]	0	0.00	0.00	1.00	0.84	0.84	0.68	0.73	0.53
[5,]	0	0.00	0.00	0.00	1.00	0.84	0.69	0.64	0.76
[6,]	0	0.00	0.00	0.00	0.00	1.00	0.83	0.73	0.76
[7,]	0	0.00	0.00	0.00	0.00	0.00	1.00	0.87	0.84
[8,]	0	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.78
[9,]	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

```
Number of observations (pairs):

     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,]   68   66   64   50   22    2    0    0    0
[2,]    0   87   83   69   39   16    0    0    0
[3,]    0    0  121  105   73   50   27    3    0
[4,]    0    0    0  127   93   70   45   19    4
[5,]    0    0    0    0  111   85   61   33   13
[6,]    0    0    0    0    0  102   77   47   25
[7,]    0    0    0    0    0    0   85   54   33
[8,]    0    0    0    0    0    0    0   66   42
[9,]    0    0    0    0    0    0    0    0   48
```

How to build models for correlation?

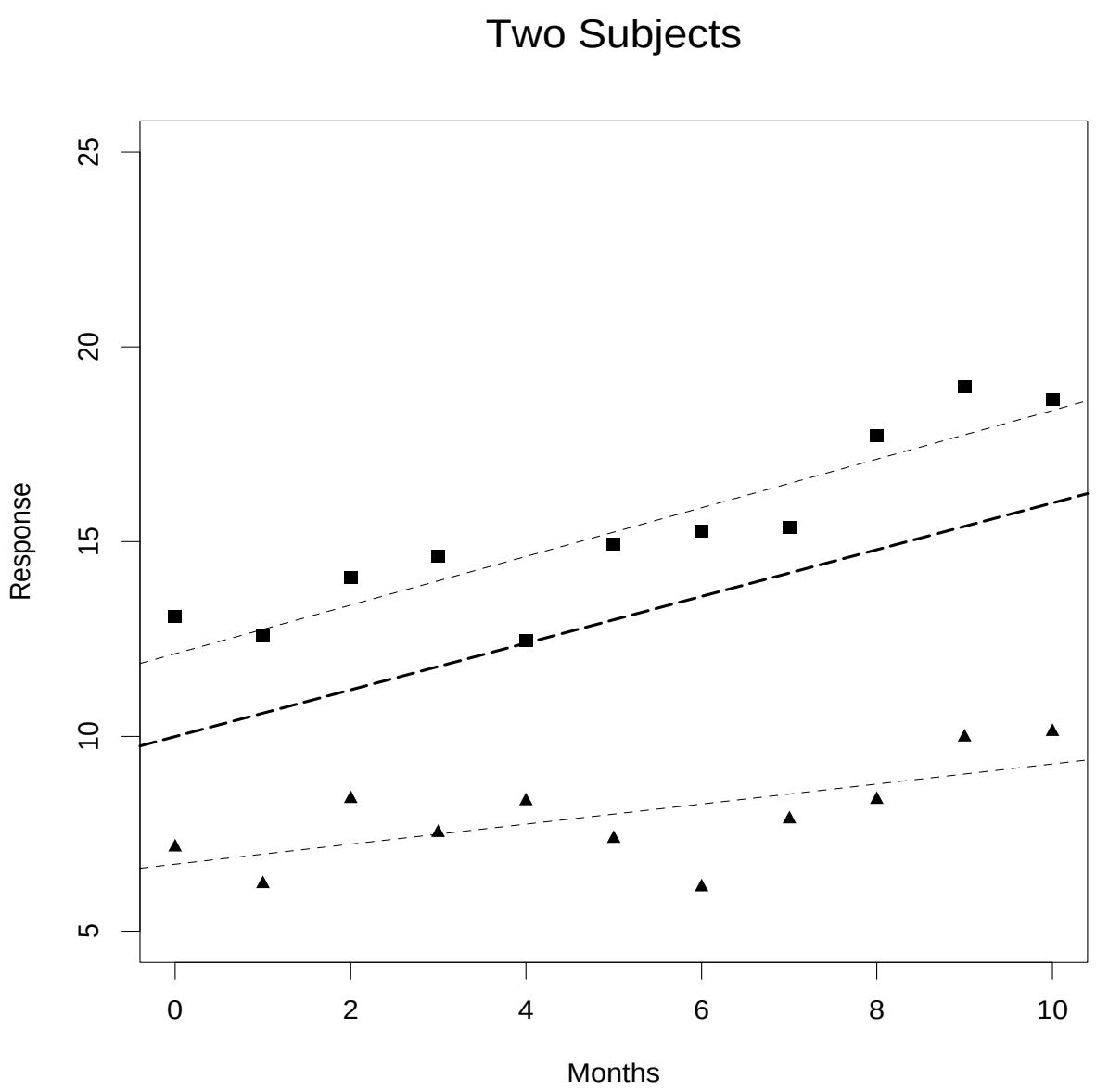
- Mixed models
  - "random effects"
  - within-subject similarity due to sharing trajectory
- Serial correlation
  - close in time implies strong similarity
  - correlation decreases as time separation increases

Linear Mixed Model

- **Regression model:**
  mean response as a function of covariates.
  "systematic variation"

- **Random effects:**
  variation from subject-to-subject in trajectory.
  "random between-subject variation"

- **Within-subject variation:**
  variation of individual observations over time
  "random within-subject variation"


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

Levels of Analysis

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

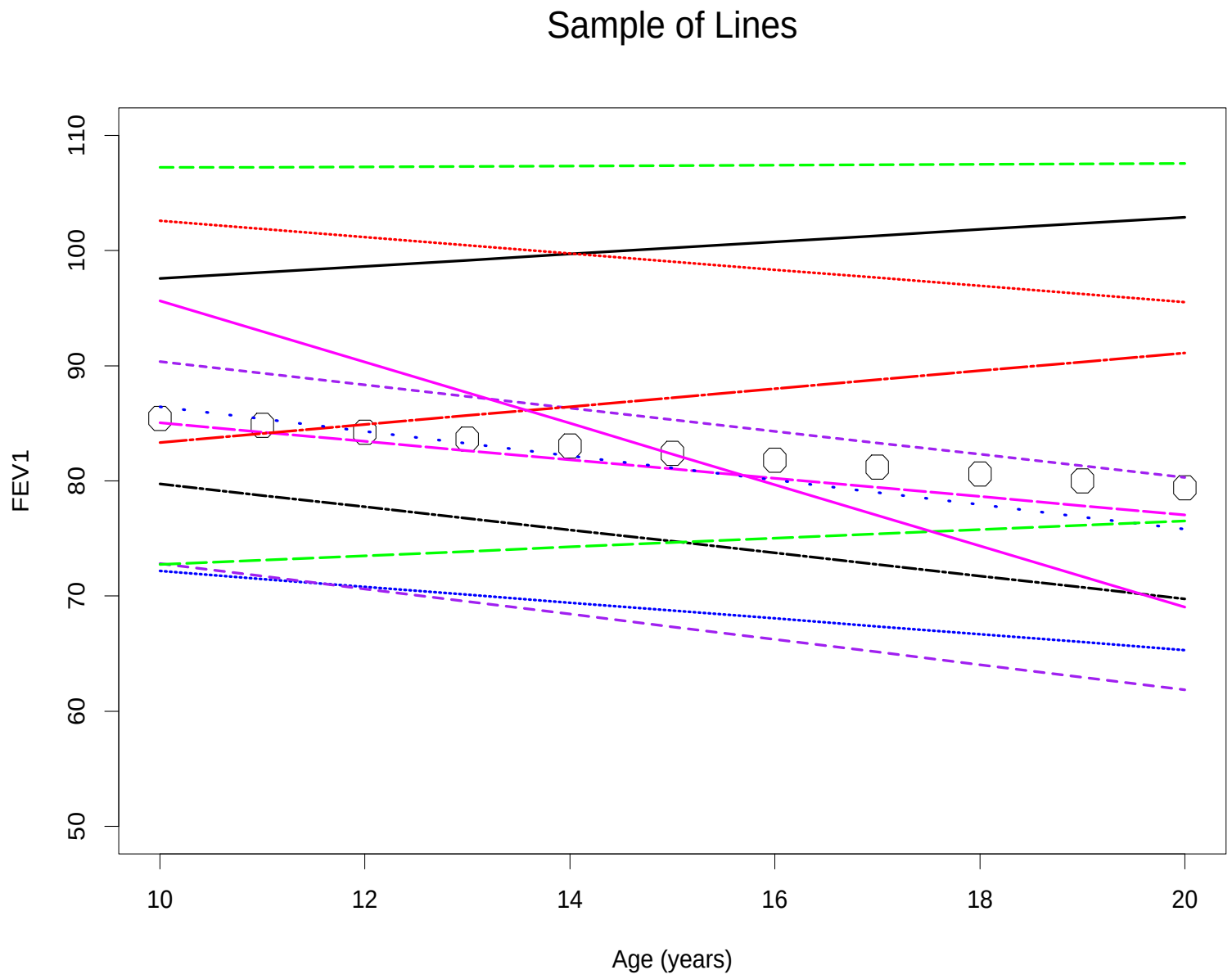
$$
\begin{aligned}
Y_{ij} &= \beta_{0,i} + \beta_{1,i} \cdot t_{ij} + e_{ij} \\
e_{ij} &\sim \mathcal{N}(0, \sigma^2) \\
E[\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{\beta}_i] &= \beta_{0,i} + \beta_{1,i} \cdot t_{ij} \\
&= [1, \texttt{time}_{ij}] \begin{bmatrix} \beta_{0,i} \\ \beta_{1,i} \end{bmatrix}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

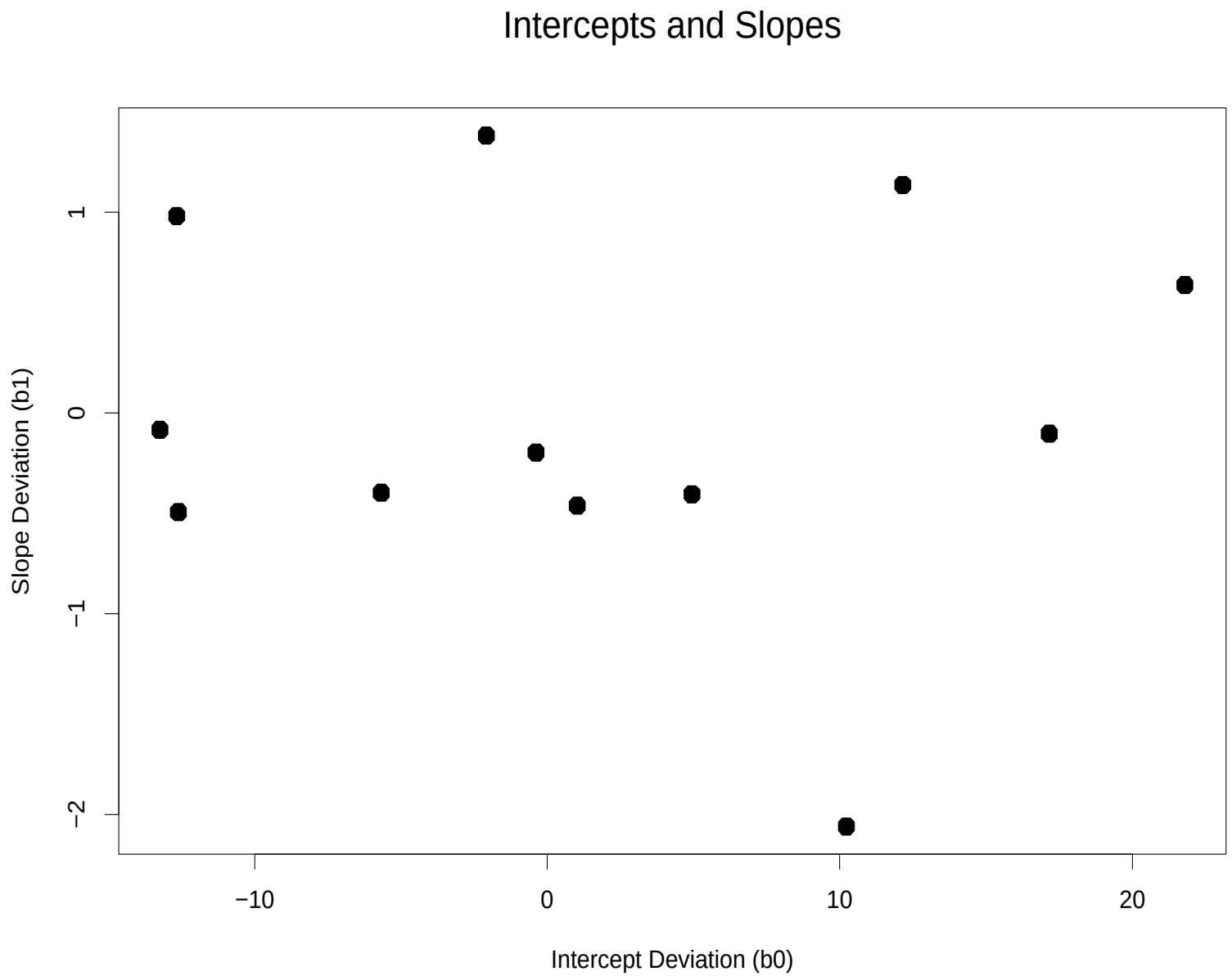
Levels of Analysis

- 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
FEV1
110
100
90
80
70
60
50
10
12
14
16
18
20
Age (years)


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

Levels of Analysis

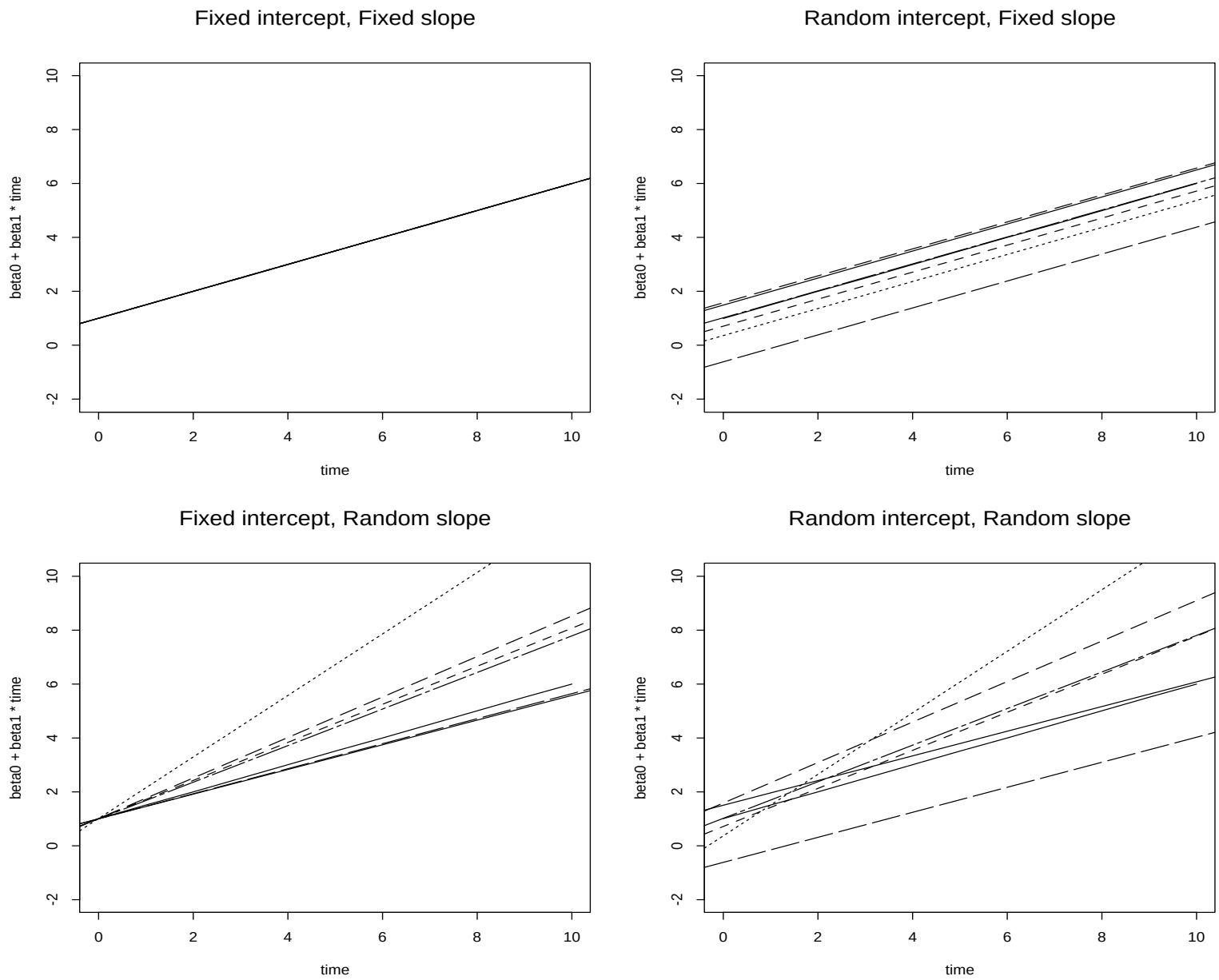
- Next we consider the distribution of **patterns (parameters)** among subjects:

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

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

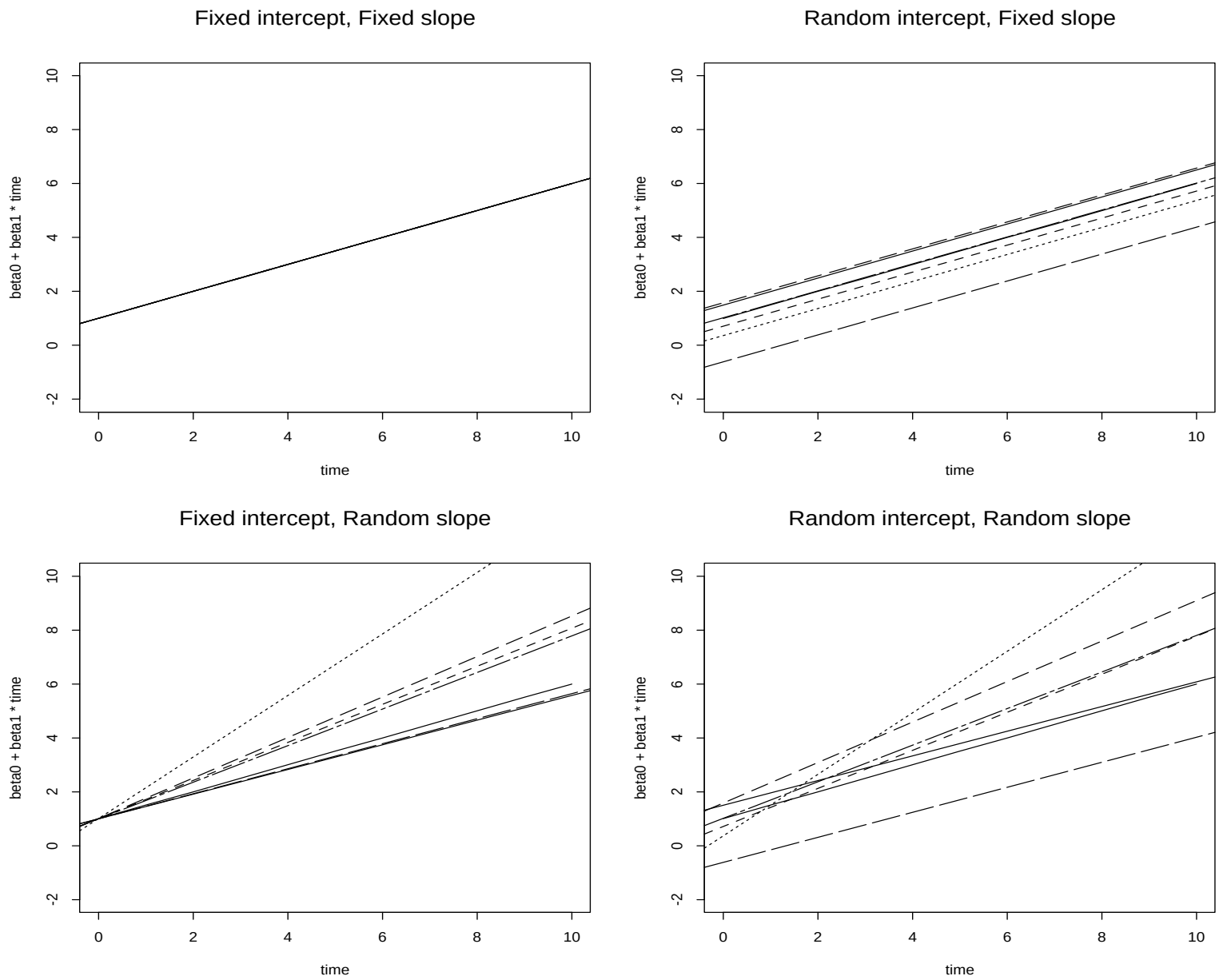
- We can use the idea of random effects to allow different types of between-subject heterogeneity:
- The magnitude of heterogeneity is characterized by $\boldsymbol{D}$:

$$\boldsymbol{b}_i = \begin{bmatrix} b_{0,i} \\ b_{1,i} \end{bmatrix}$$

$$\text{var}(\boldsymbol{b}_i) = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}$$

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

- No random effects:

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

- Random intercepts:

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

- Random intercepts and slopes:

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

Mixed Models and Covariances/Correlation

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

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

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

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

Mixed Models and Covariances/Correlation

- Random Intercept Model

$$\text{corr}(Y_{ij}, Y_{ik}) = \frac{D_{11}}{\sqrt{D_{11}+\sigma^2}\sqrt{D_{11}+\sigma^2}}$$

$$= \frac{D_{11}}{D_{11}+\sigma^2} = \frac{\textbf{between var}}{\textbf{between var} + \textbf{within var}}$$

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

Mixed Models and Covariances/Correlation

- Random Intercept and Slope Model

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

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

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

Mixed Models and Covariances/Correlation

- 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 $\text{cov}(e_{ij}, e_{ik}) = 0$.

Linear Mixed Model

- **Regression model:**
  mean response as a function of covariates.
  "systematic variation"

- **Random effects:**
  variation from subject-to-subject in trajectory.
  "random between-subject variation"

- **Within-subject variation:**
  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

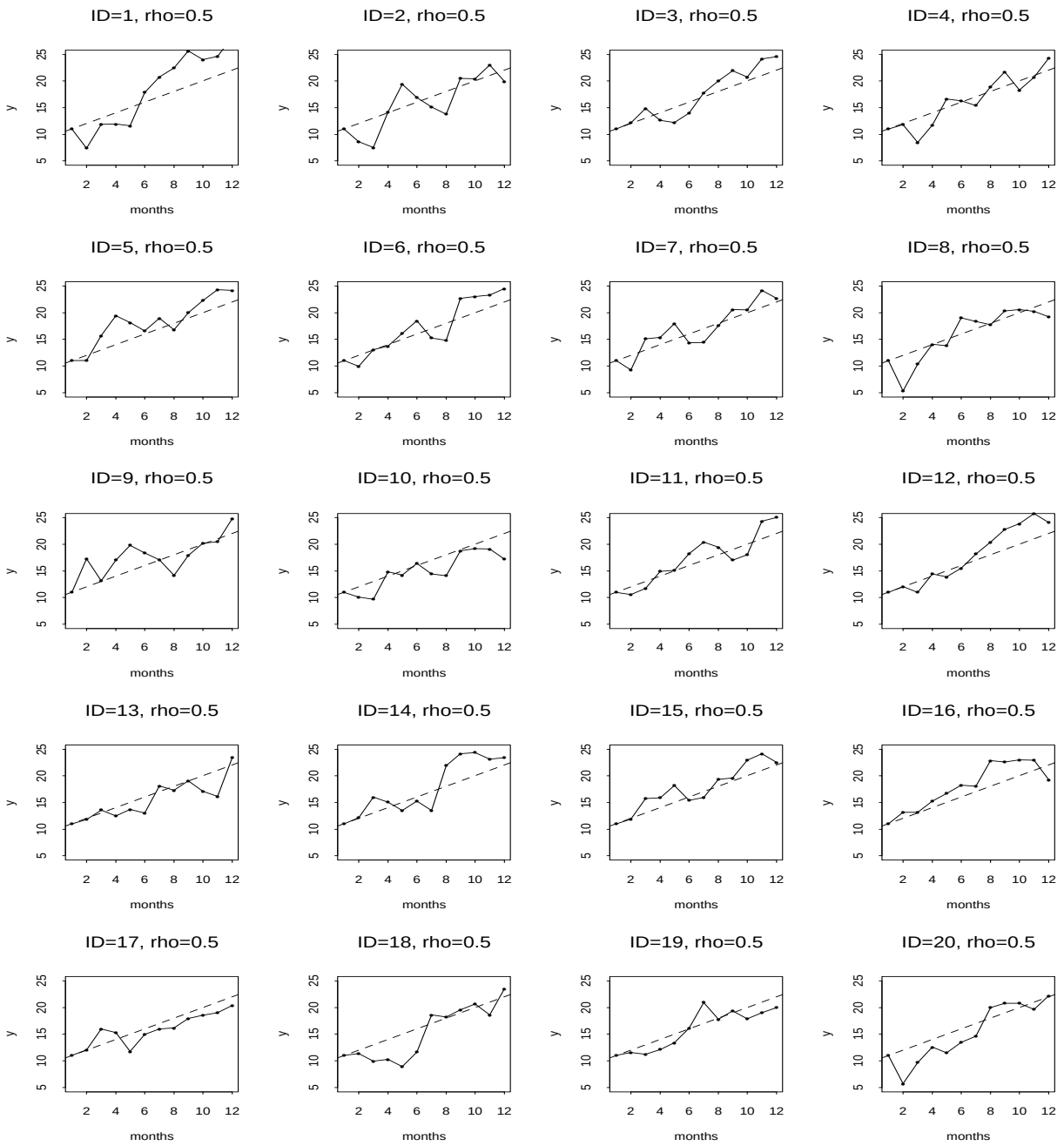
Define

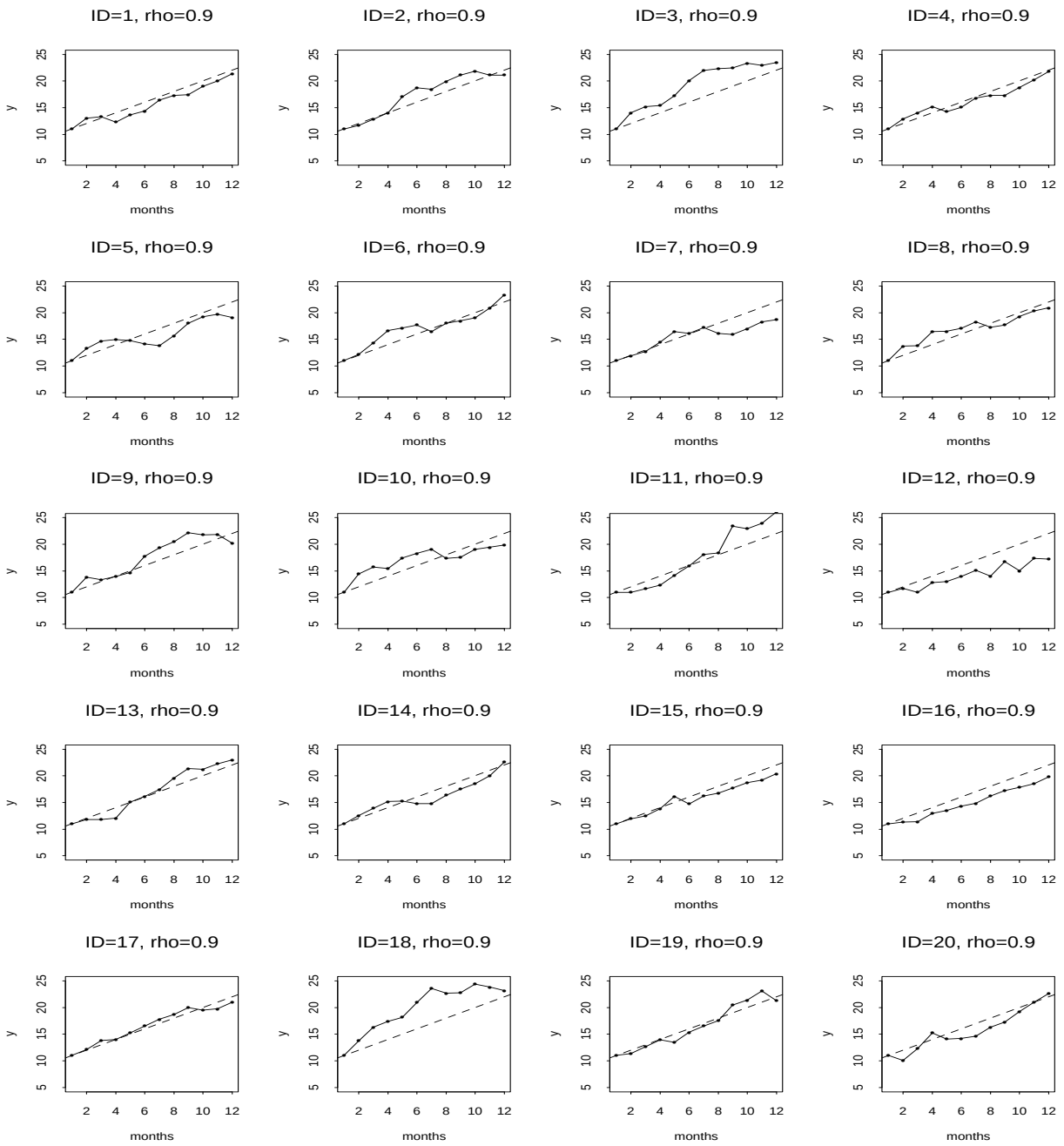
$$e_{ij} = \rho \cdot e_{ij-1} + \epsilon_{ij}$$

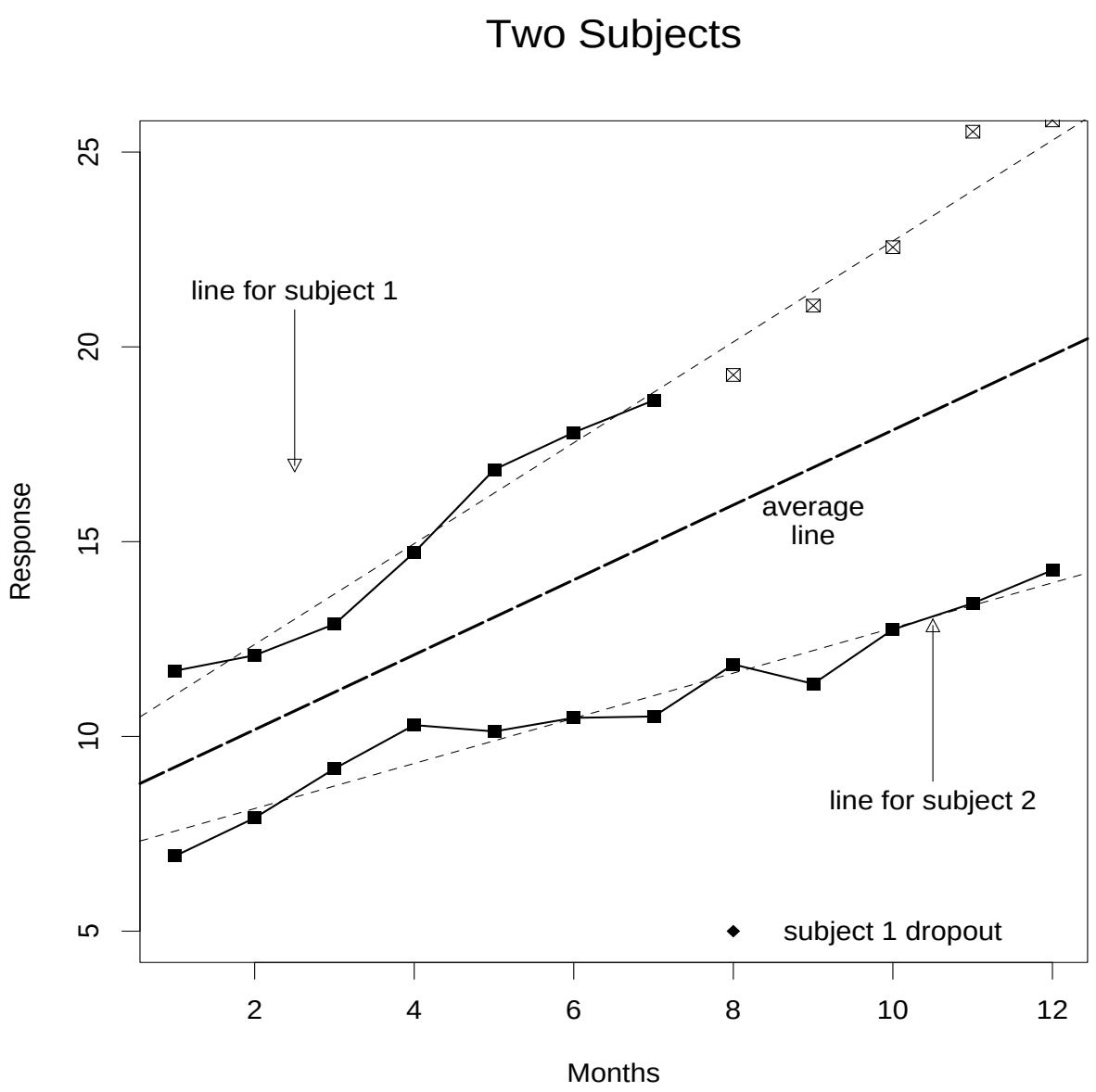
$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2(1-\rho^2))$$
$$\epsilon_{i0} \sim \mathcal{N}(0, \sigma^2)$$

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
months
5 10 15 20 25
2 4 6 8 10 12


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

Linear Mixed Model

- **Regression model**:
  mean response as a function of covariates.
  "systematic variation"

- **Random effects**:
  variation from subject-to-subject in trajectory.
  "random between-subject variation"

- **Within-subject variation**:
  variation of individual observations over time
  "random within-subject variation"