Analysis of Longitudinal Data

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

Session One Outline

- Examples of longitudinal data
- Scientific motivation
  - Opportunities
  - Issues
- Time scales
  - Cross-sectional contrasts
  - Longitudinal contrasts
- Exploratory data analysis
  - between- and within-person variation
  - correlation / covariance

Longitudinal Data Analysis

INTRODUCTION

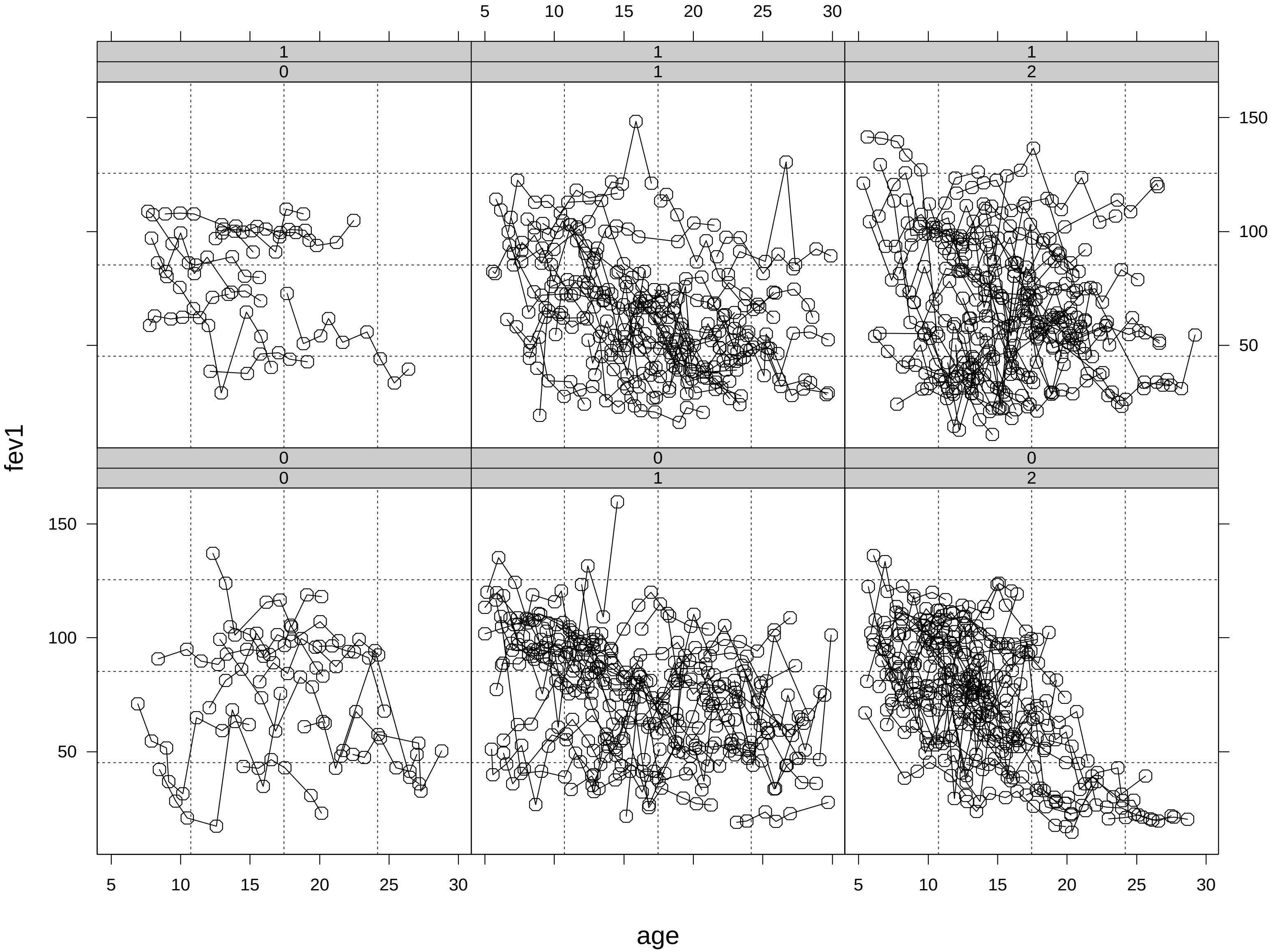
to

EXAMPLES AND ISSUES

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?



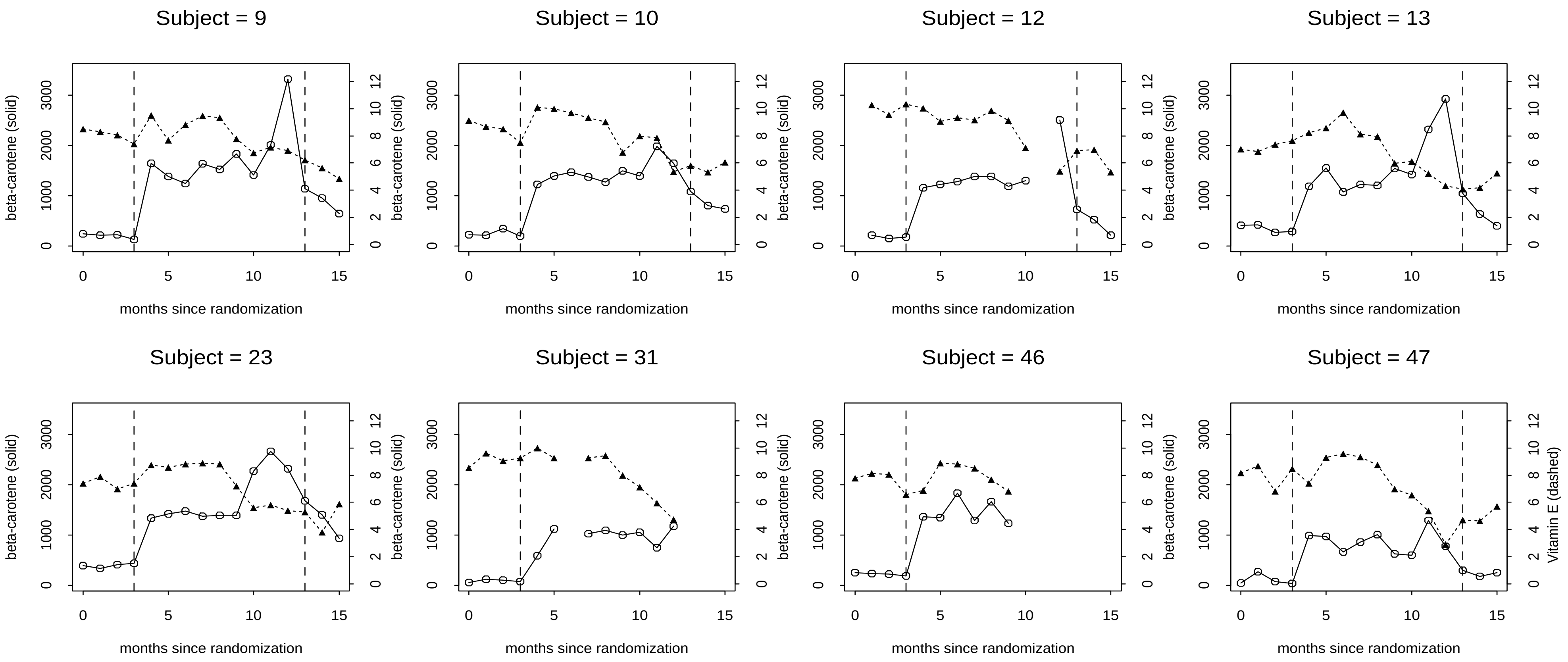
fev1
age
1
0
1
1
1
2
0
0
0
1
0
2
5
10
15
20
25
30
50
100
150

Continuous Longitudinal Data

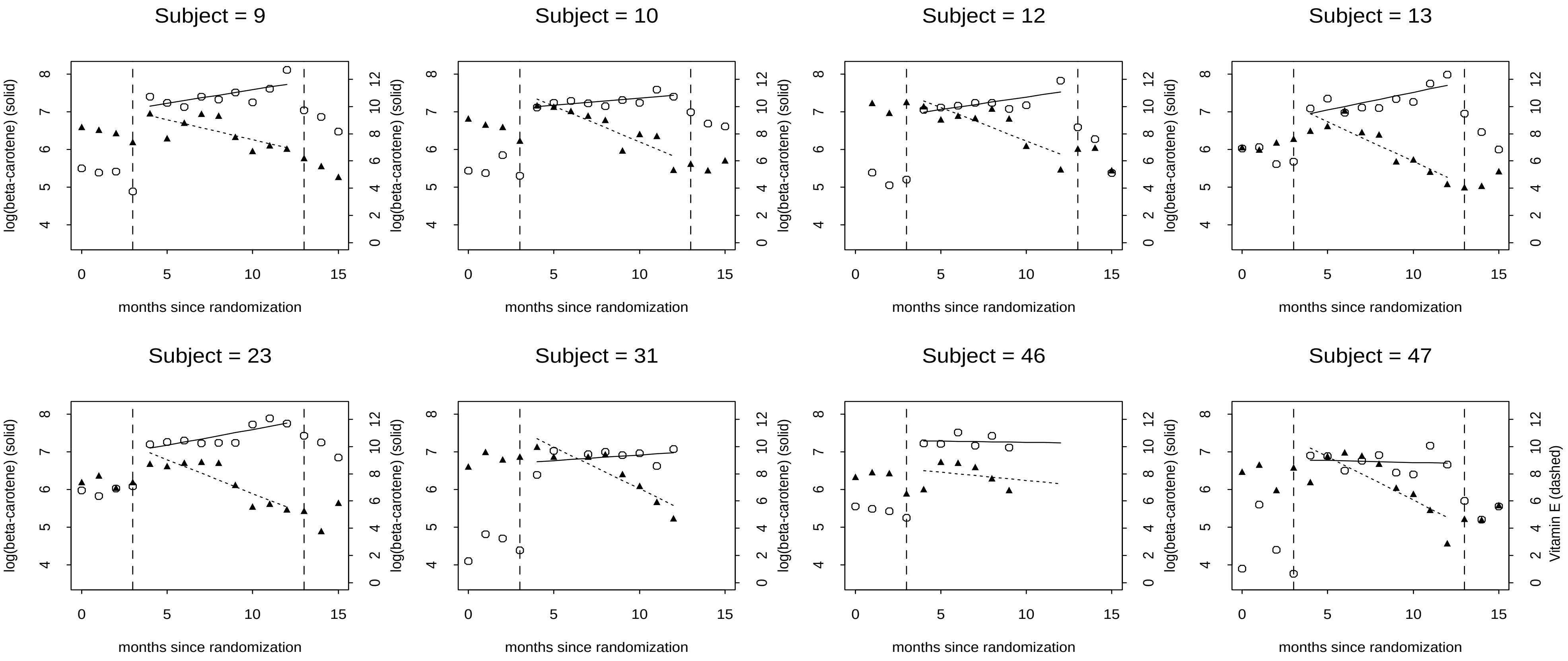
Example 2: Beta-carotene and vitamin E

- 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
5
10
15
0
1000
2000
3000
0
2
4
6
8
10
12

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
0
5
10
15
4
5
6
7
8
0
2
4
6
8
10
12

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?

Vol. 24, No. 2 UTILIZATION OF PEDIATRIC AMBULATORY CARE

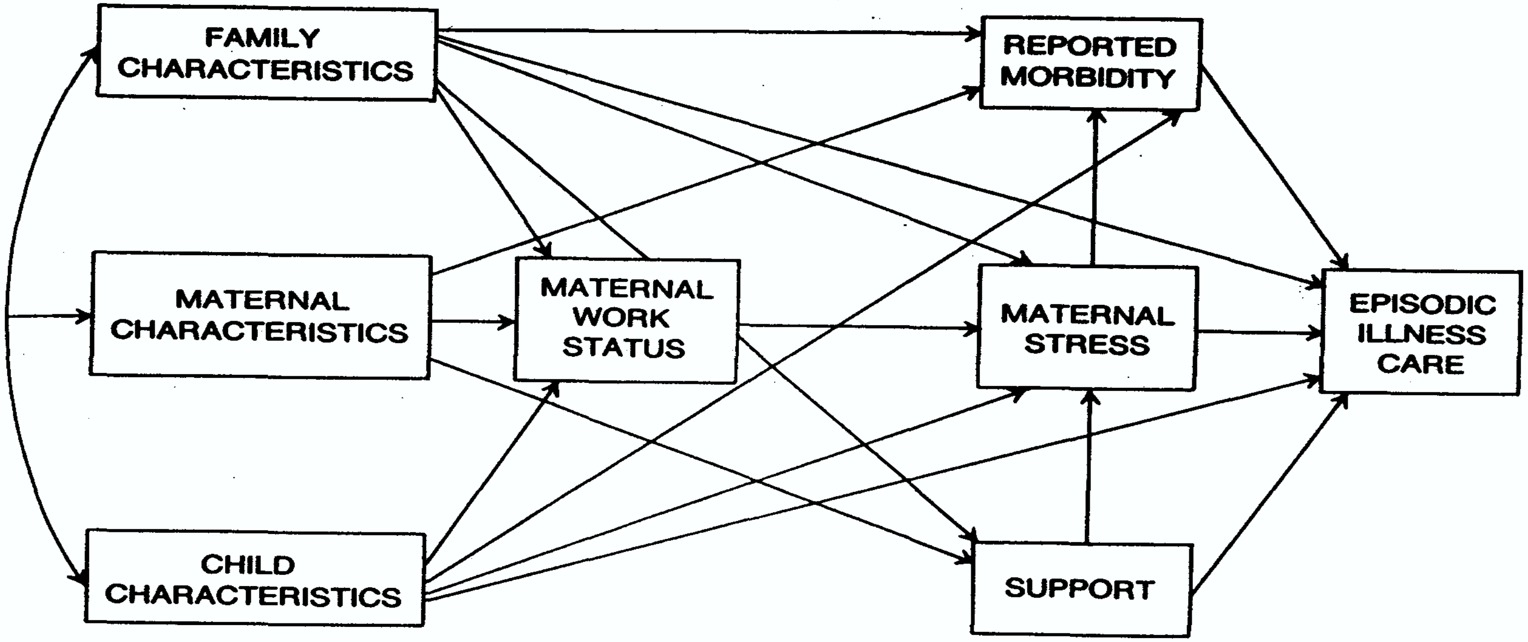
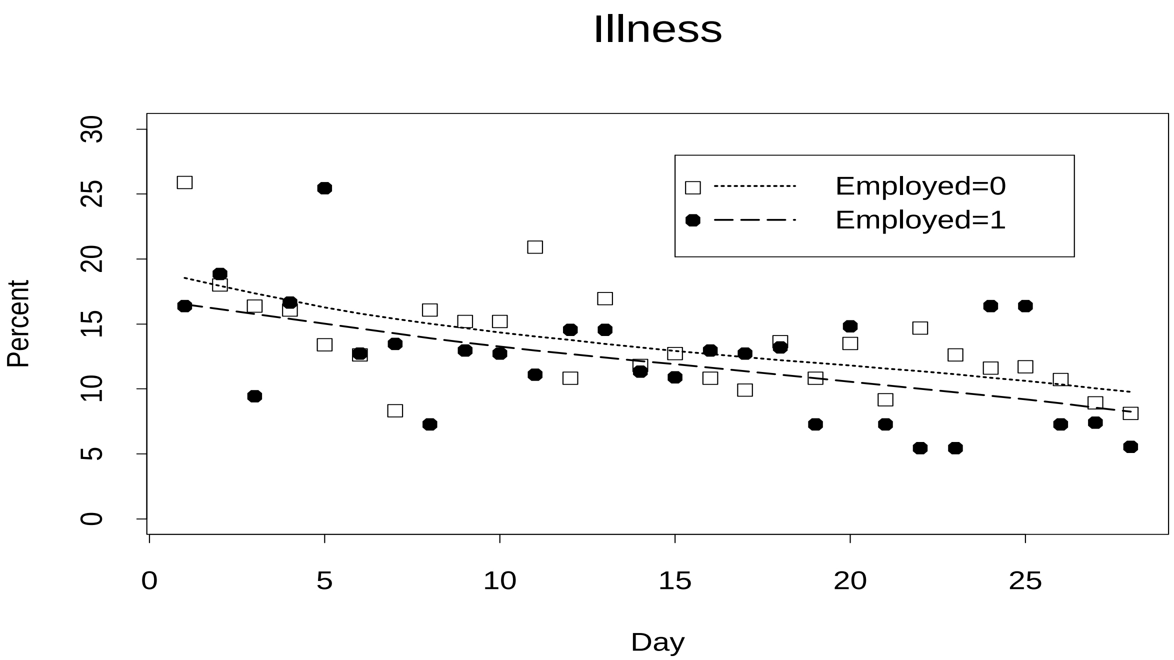
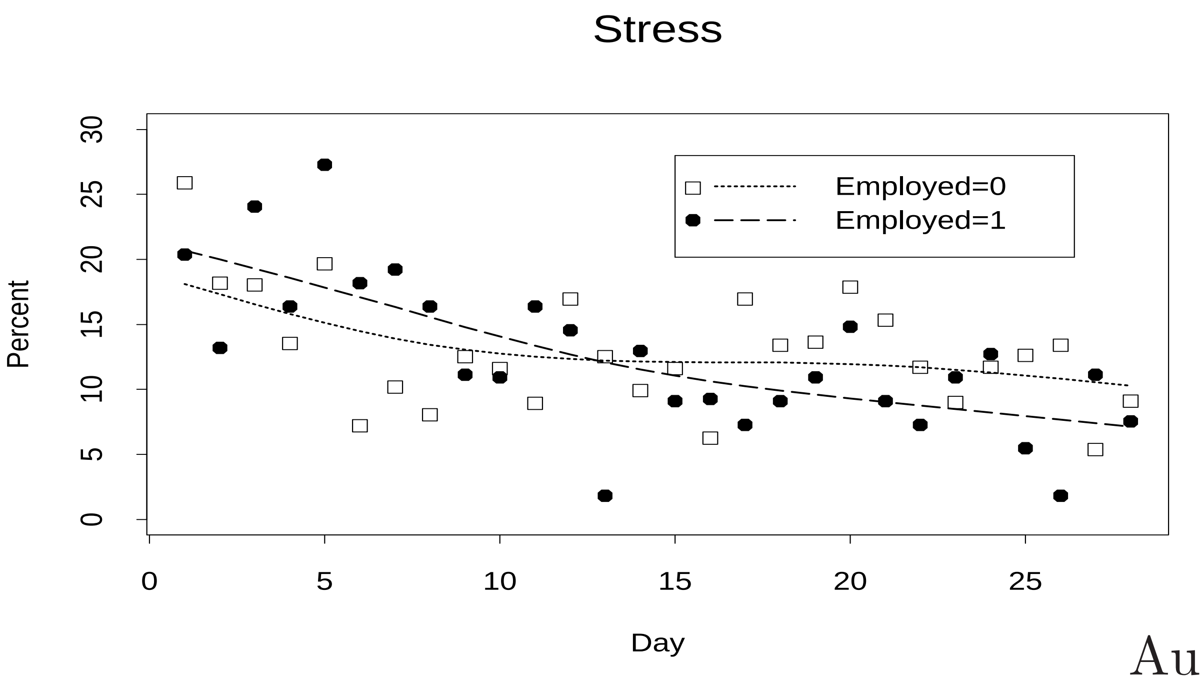
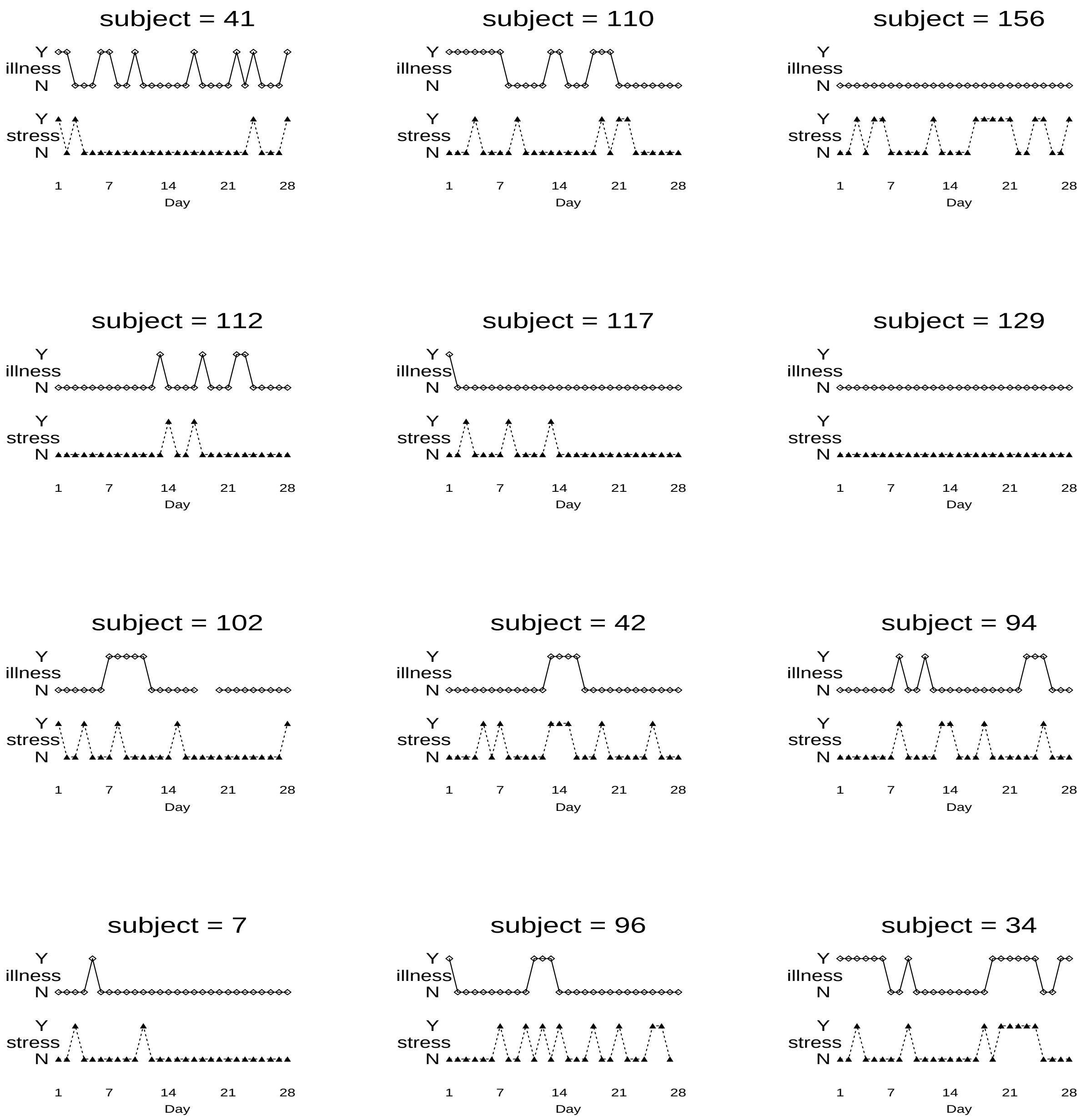


FIG. 1. Determinants of episodic illness care utilization.


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


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


subject = 41
subject = 110
subject = 156
subject = 112
subject = 117
subject = 129
subject = 102
subject = 42
subject = 94
subject = 7
subject = 96
subject = 34
Y
illness
N
Y
stress
N
1
7
14
21
28
Day

Continuous Longitudinal Data

Example 4: PANSS Data

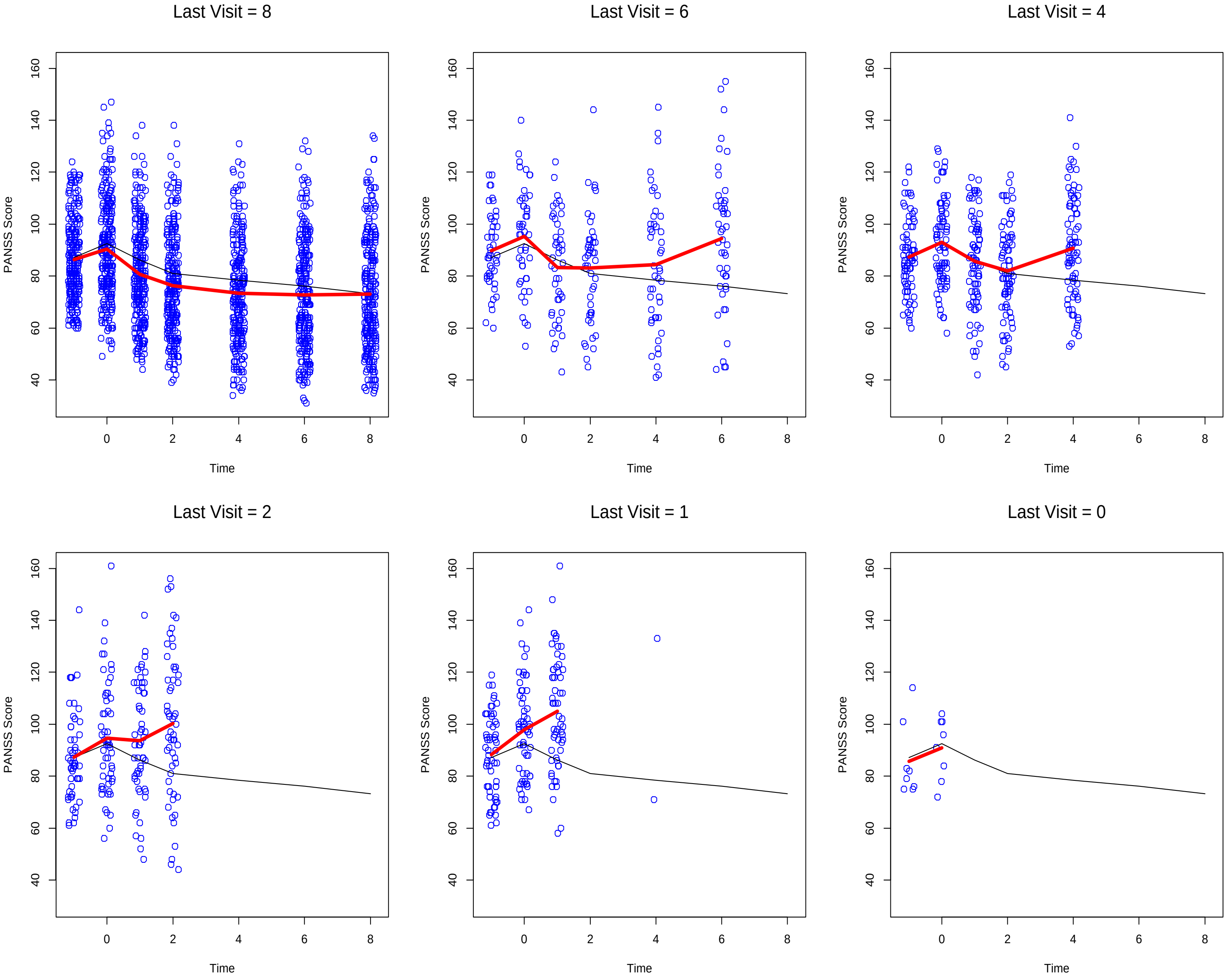
- PANSS is a standard symptom assessment for schizophrenic patients. This study compares different doses of a new agent to a standard agent and to placebo.
- primary outcome: PANSS
- covariates: treatment, time.
- **Q**: What's the best treatment?

Schizophrenia Treatment Trial

- Reasons for dropout:

Abnormal lab result	4
Adverse experience	26
Inadequate response	183
Inter-current illness	3
Lost to follow-up	3
Uncooperative	25
Withdrew consent	19
Other	7

- This combines the 6 treatment arms


Last Visit = 8
Last Visit = 6
Last Visit = 4
Last Visit = 2
Last Visit = 1
Last Visit = 0
PANSS Score
Time
40
60
80
100
120
140
160
0
2
4
6
8

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.

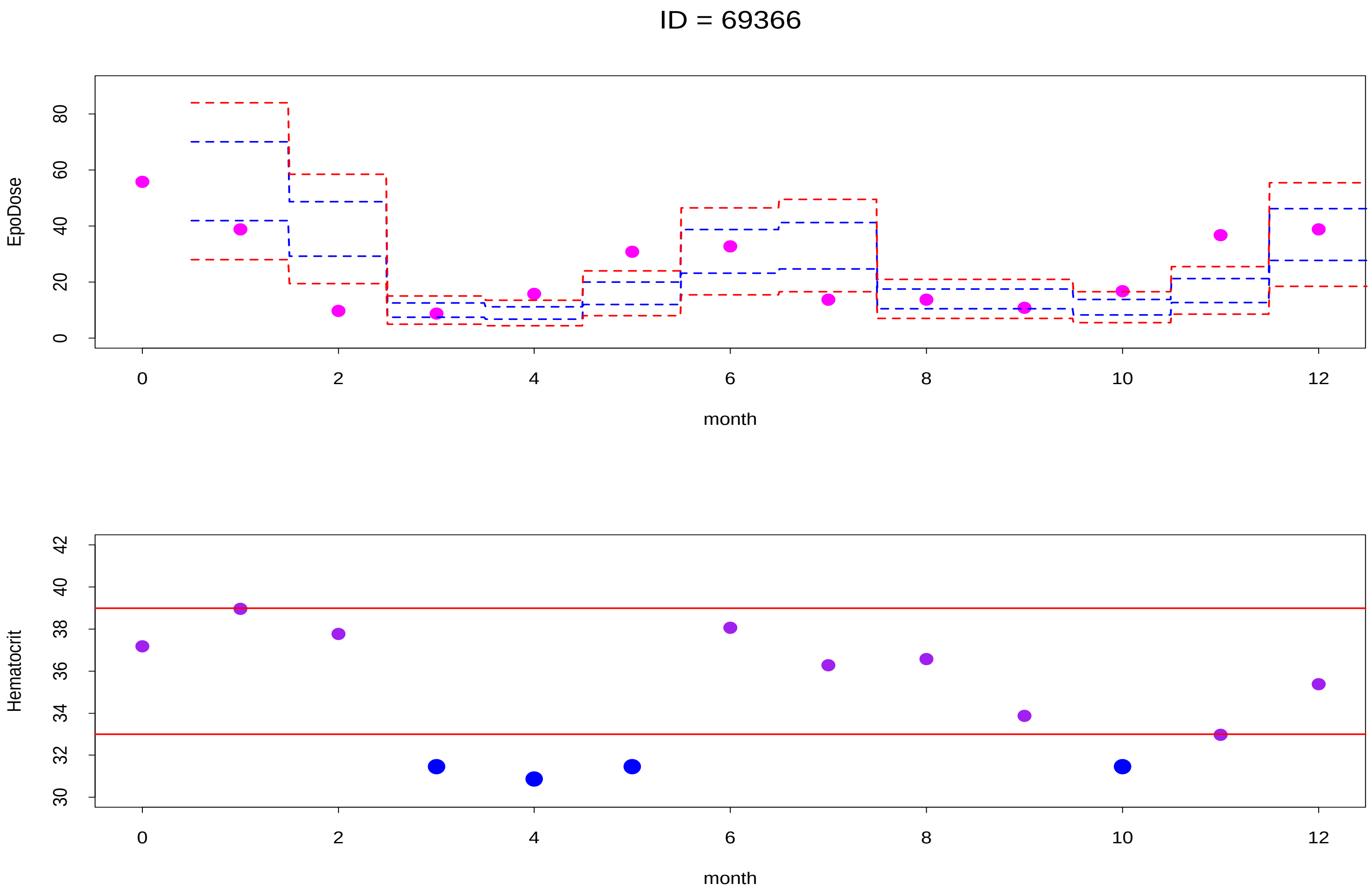
2. **Analysis of correlated data**

- 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.

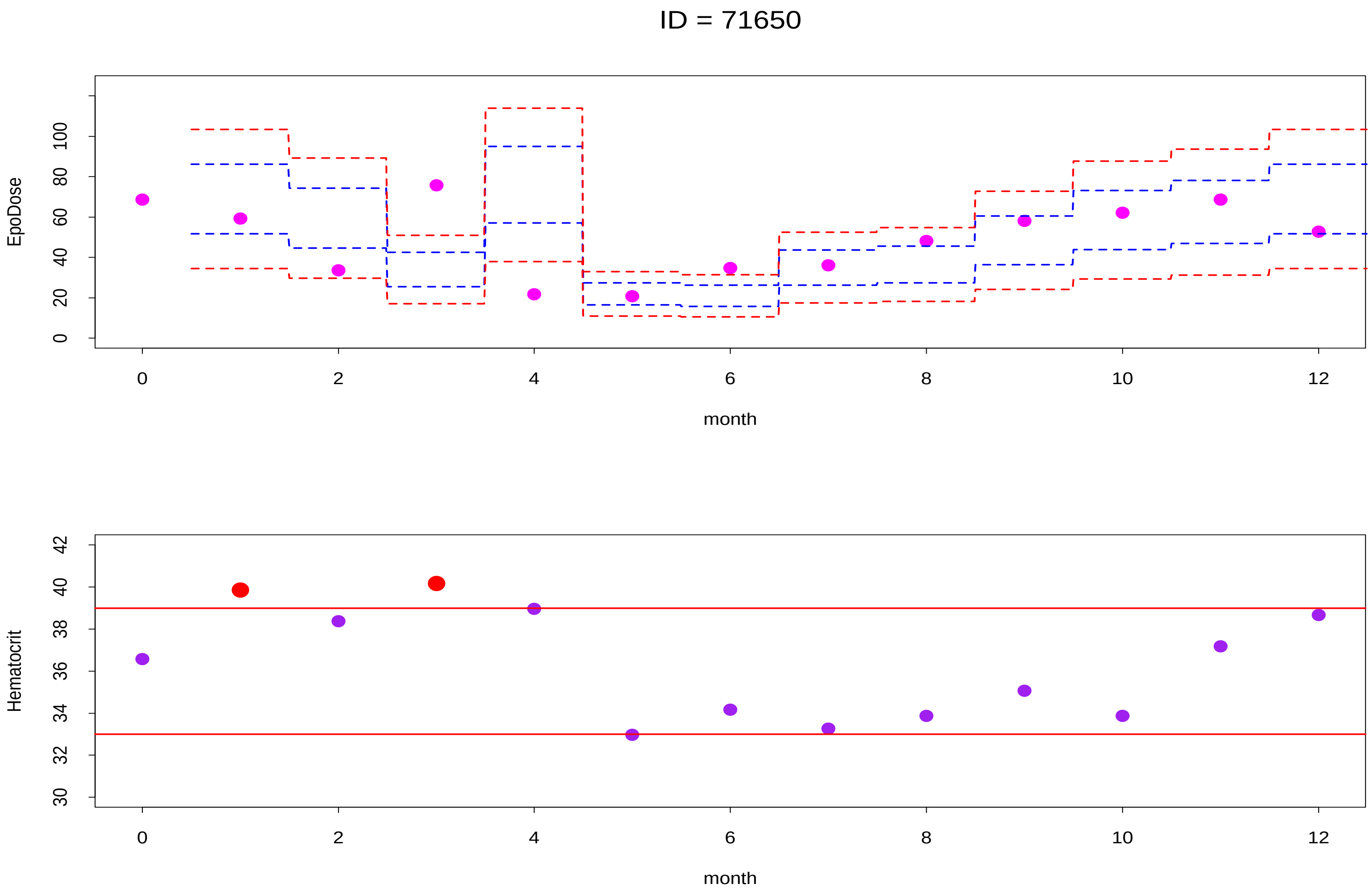
3. **Time-varying covariates**

- 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?

USRDS Dialysis Data



USRDS Dialysis Data



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.

$\star \star \star$ **Today's focus**: 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?
   - **Change**
2. Is the time course different for males and females?
   - **Group Differences**
3. Is the time course different for F508 homozygous subjects ?
   - **Group Differences**

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

Next: some exploratory data analysis (EDA)

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)
```

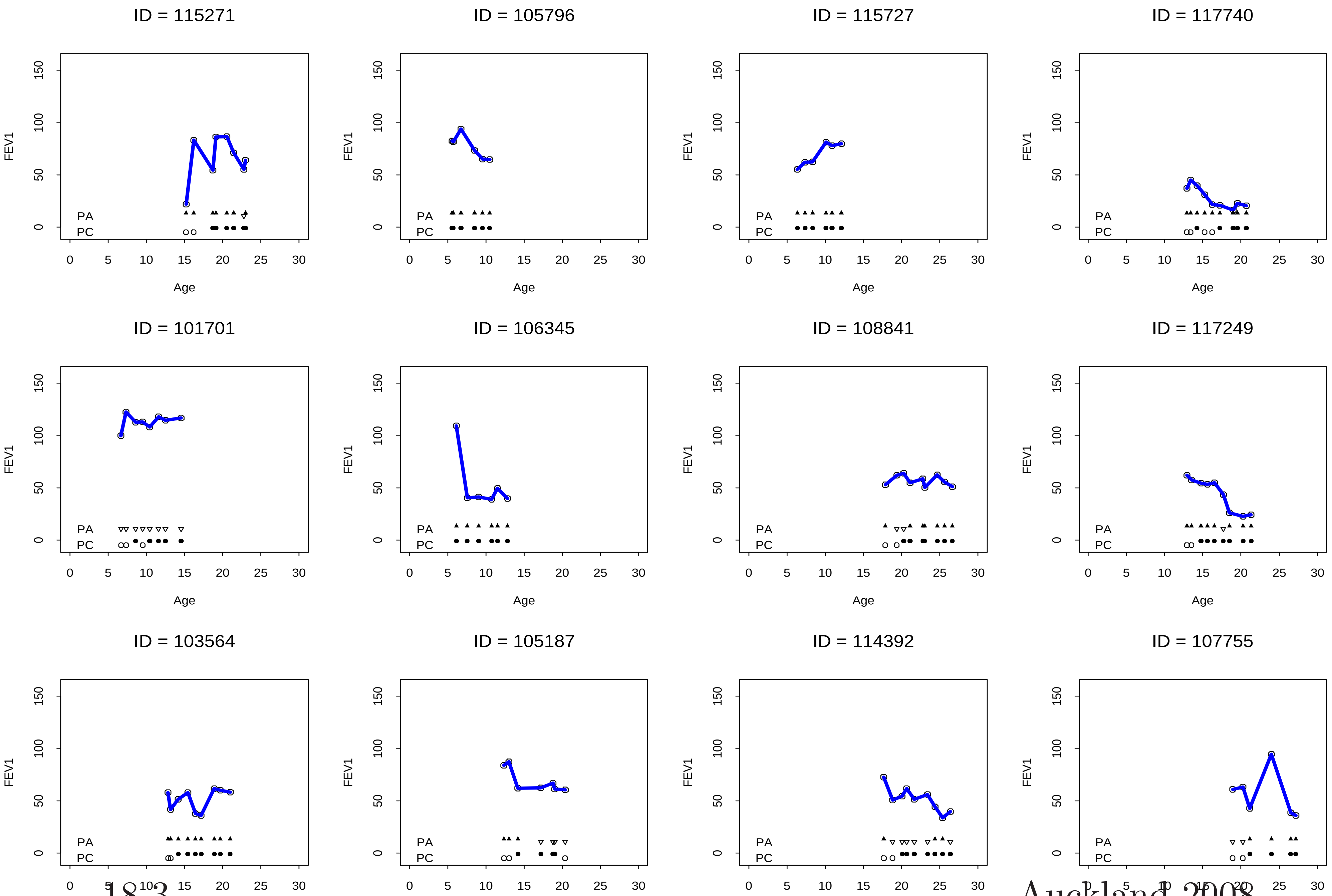
```
100073 113.8   8.452 2 3 1 2
100073  98.18  8.783 2 3 1 2
100073  98.73  9.785 2 3 1 2
100073 101.79 10.538 2 3 1 2
100073  98.04 12.329 2 3 1 2
100073  94.32 13.306 2 3 1 2
100073  95.48 14.418 2 3 1 2
100111  96.85 12.515 2 0 3 1
100111 101.05 13.103 2 0 3 2
100111 100.33 15.105 2 0 3 2
100111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

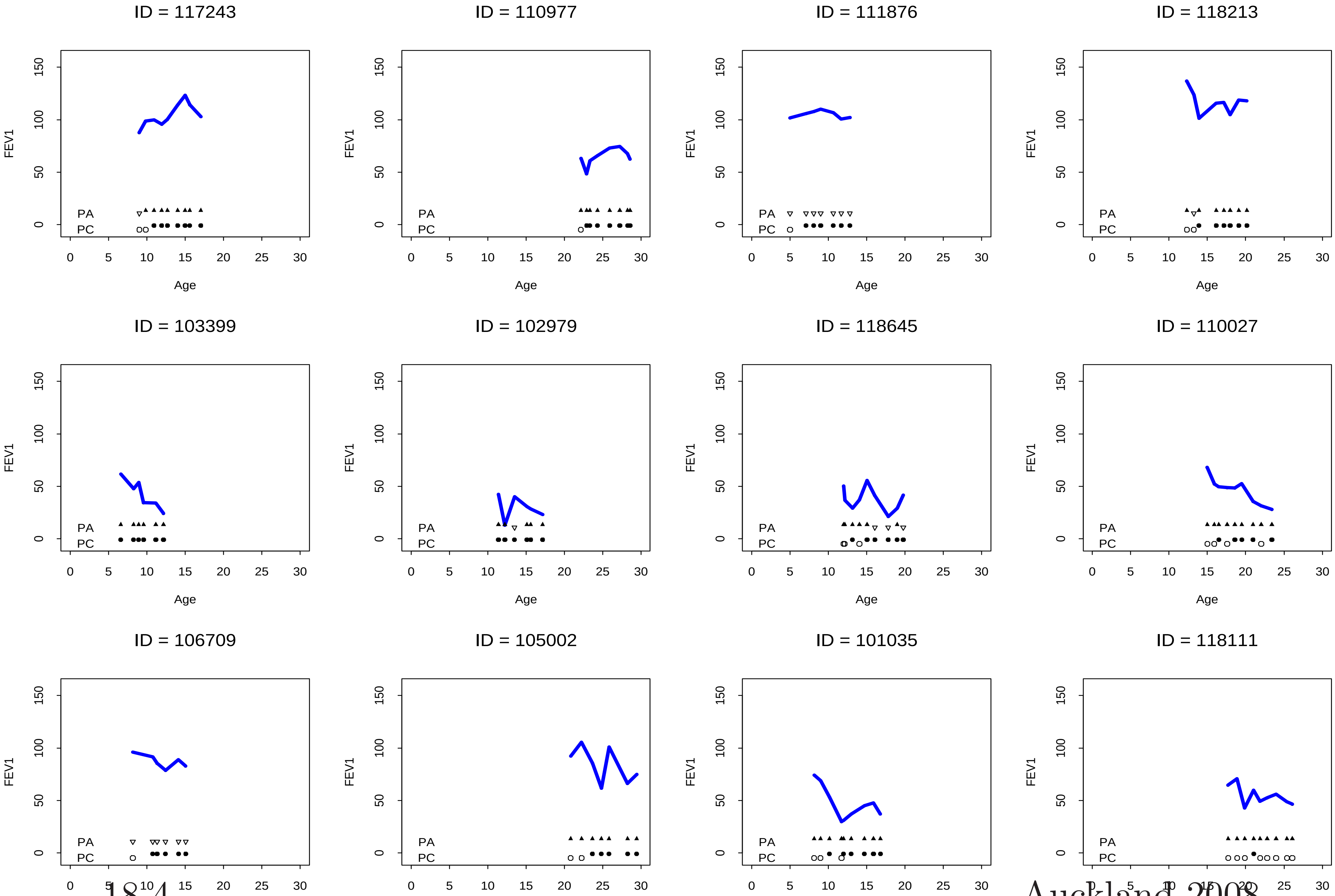
EDA: Numerical Summaries

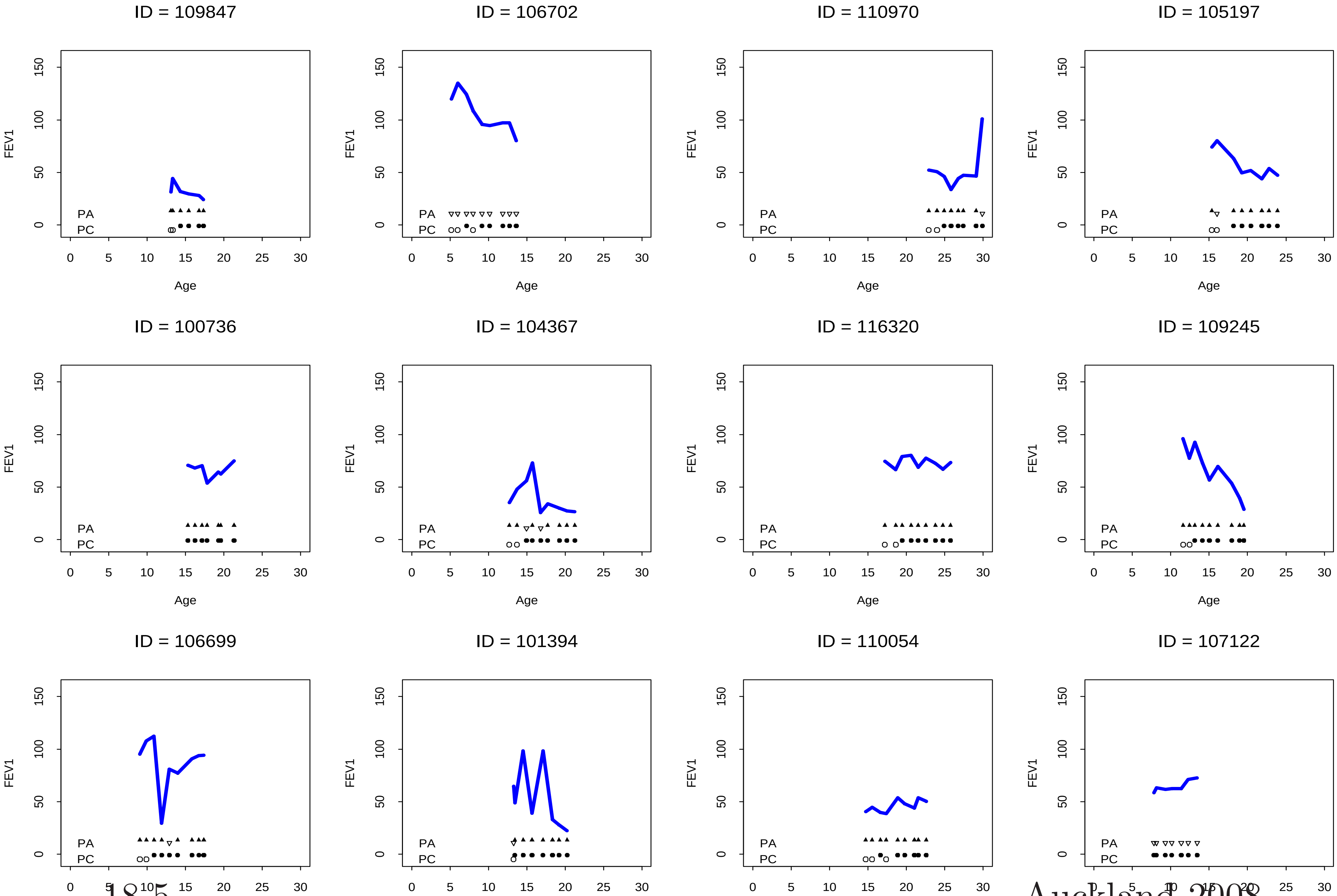
```
Total number of subjects = 200


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
```

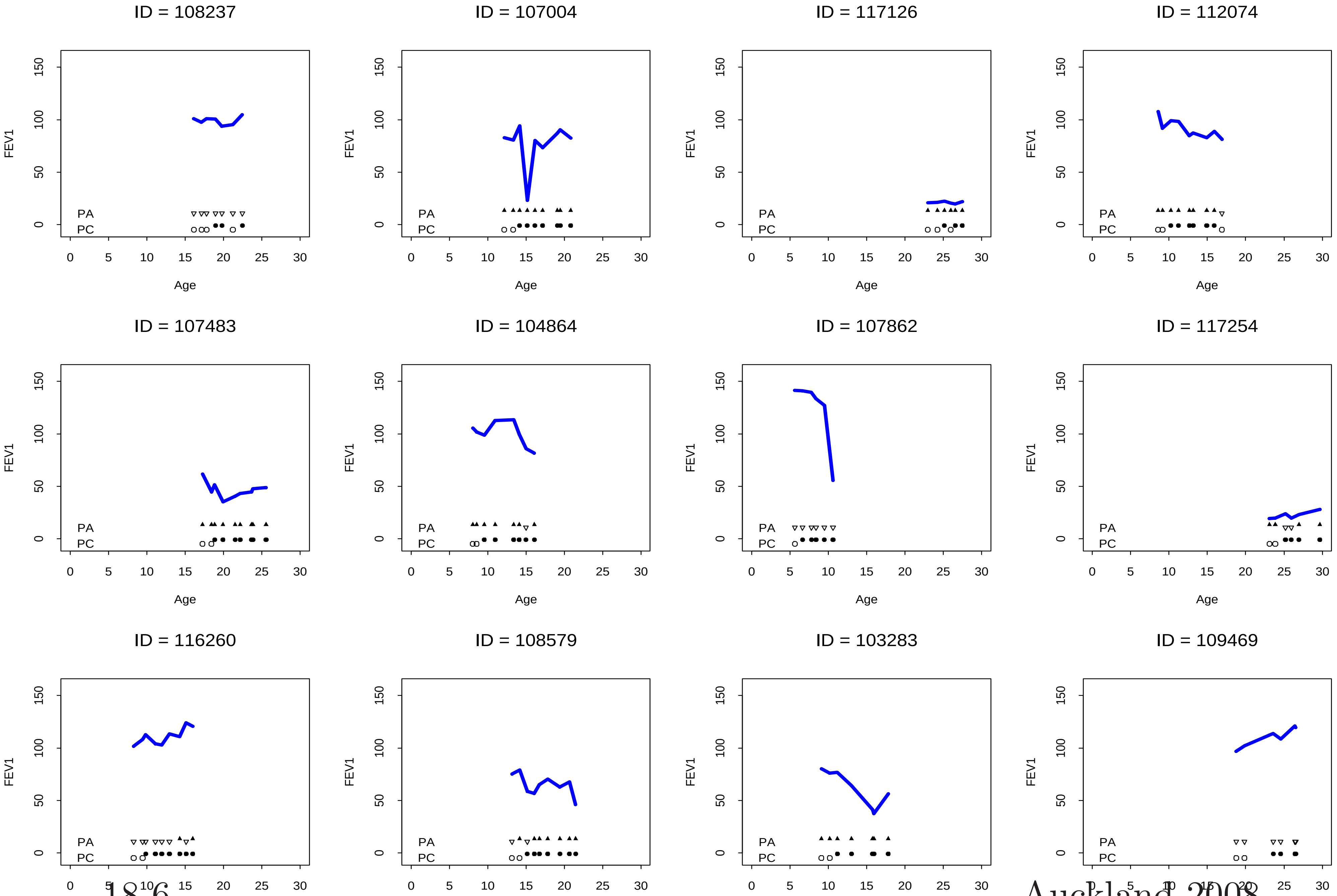
EDA: Numerical Summaries

```
Age at entry

N = 200   Median = 11.9655
Quartiles = 7.758, 15.3235

Decimal point is at the colon

    5 : 00235566678889
    6 : 0111222334555567789999
    7 : 000123444678889
    8 : 01122334556689
    9 : 00011244788
   10 : 0111113349
   11 : 2223446678
   12 : 0011122233445557788888999
   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
0
50
100
150
Age
0
5
10
15
20
25
30
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
0
50
100
150
PA
PC
Age
0
5
10
15
20
25
30


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
0
50
100
150
0
5
10
15
20
25
30

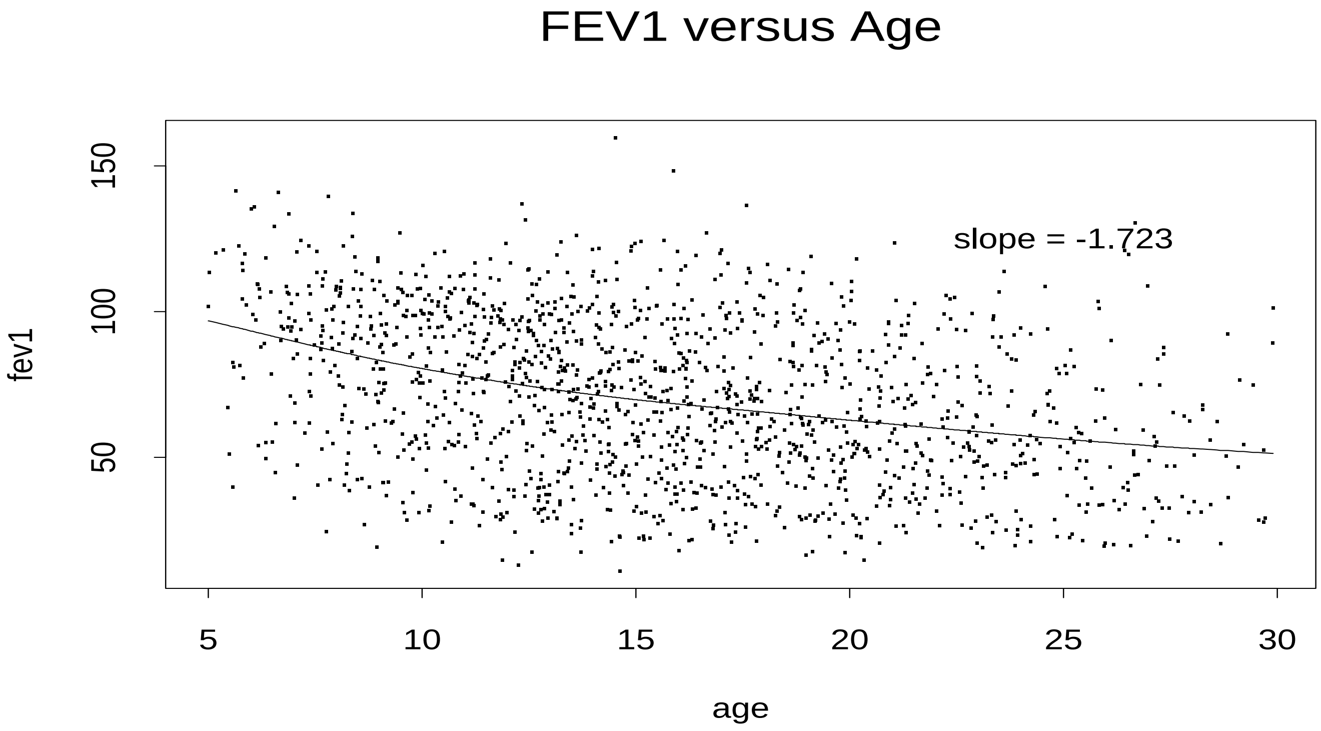

ID = 108237
ID = 107004
ID = 117126
ID = 112074
ID = 107483
ID = 104864
ID = 107862
ID = 117254
ID = 116260
ID = 108579
ID = 103283
ID = 109469
FEV1
0
50
100
150
Age
0
5
10
15
20
25
30
PA
PC

Choosing Time Scale(s)

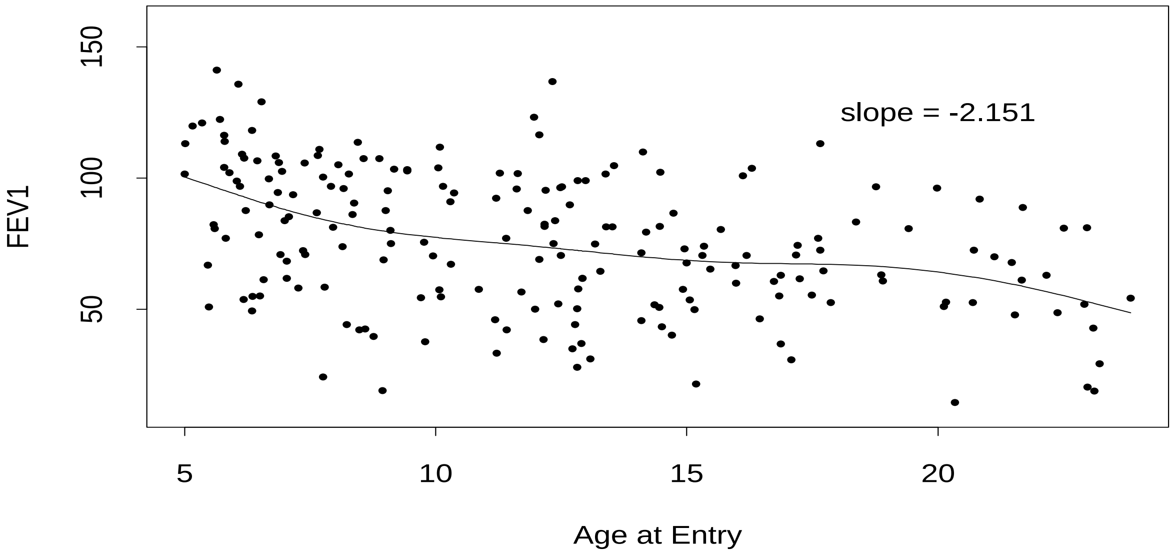
- **Age**: use $\texttt{AGE}_{ij}$ as the time variable.
  - ▷ Assumes: decline from age 10 to age 12 experienced 1981–1983 is the same as that from from age 10 to age 12 experienced 1991–1993.
  - ▷ (e.g. no **period** effects)
- **Age-since-entry**: use $\texttt{AGE}_{ij} - \texttt{AGE}_{i1}$ as the time variable.
  - ▷ Here: this is the same as the calendar year (for most subjects).
  - ▷ Assumes: decline experienced from 1991-1993 is the same for children that aged from 10 to 12 years old, and children that aged from 20 to 22 years old.
  - ▷ (e.g. no **cohort** effects)

Choosing Time Scale(s)

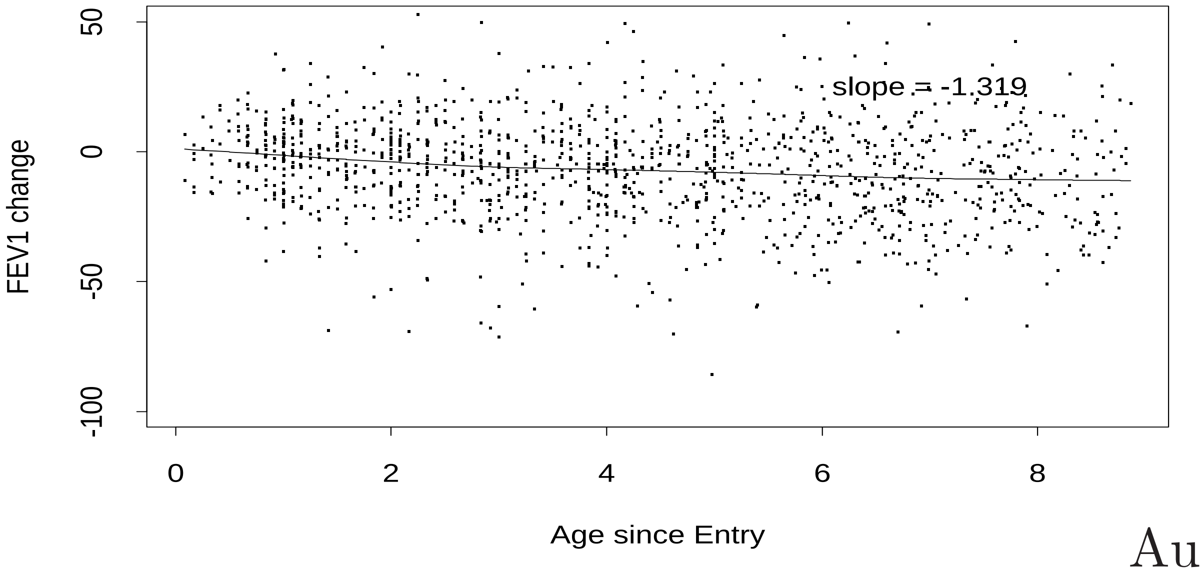
- **Age-at-entry**: use $\mathtt{AGE}_{i1}$ as the time variable.
  - ▷ Here: this is the same as CHILD AGE in the year data collection started (e.g. 1986).
  - ▷ Assumes: children may be different at entry to study, but do not change further during follow-up.
  - ▷ (e.g. no **aging** effects)


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

FEV1 versus Age-at-Entry



FEV1 Change versus 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.
- **EDA**: plot Y_{i1} versus X_{i1}.

Distinguishing Cross-sectional and Longitudinal Associations

- Longitudinal data

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

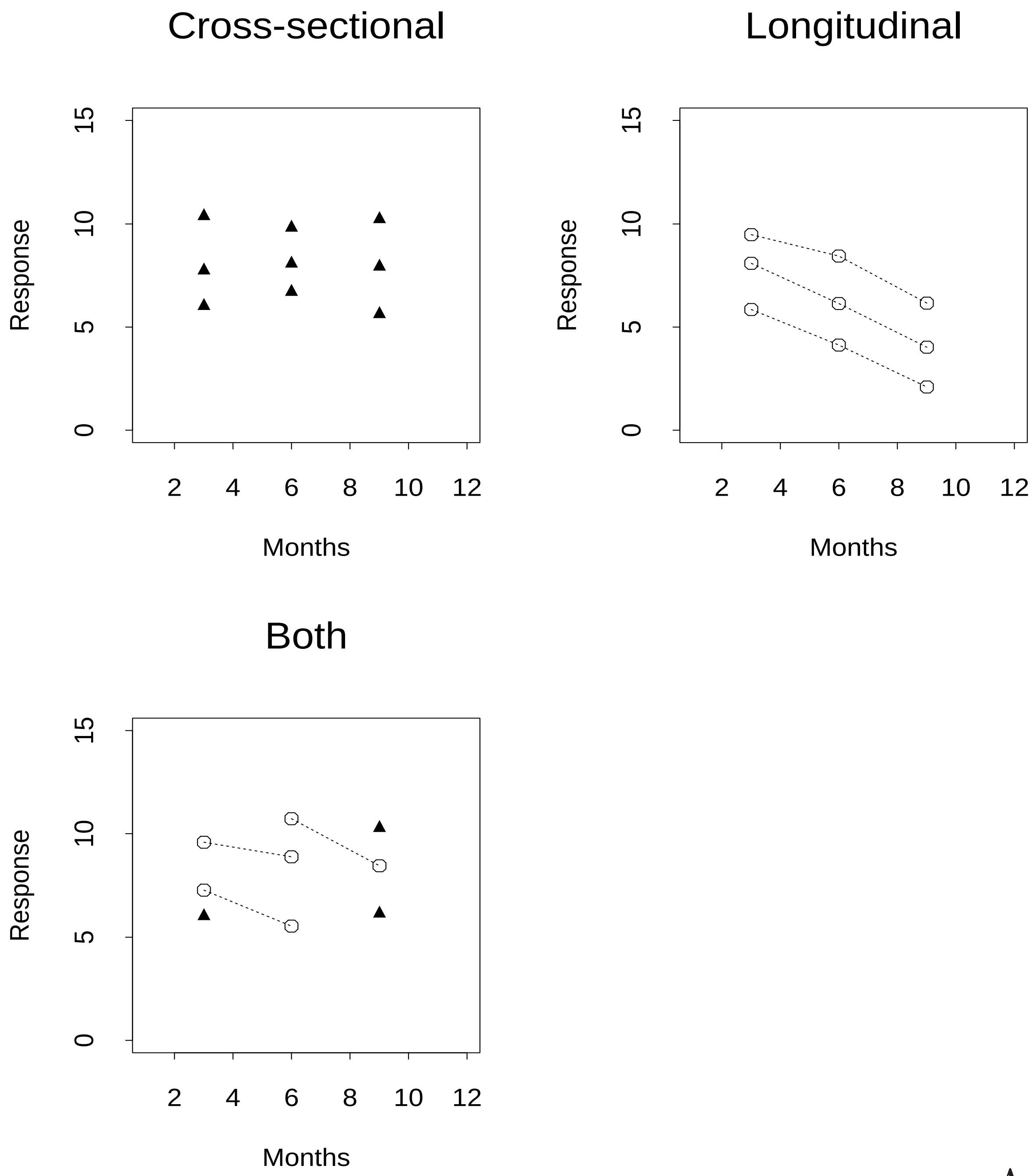
- When $j = 1$, the two equations are the same; β_C has the same cross-sectional interpretation

- Subtract equations above to obtain

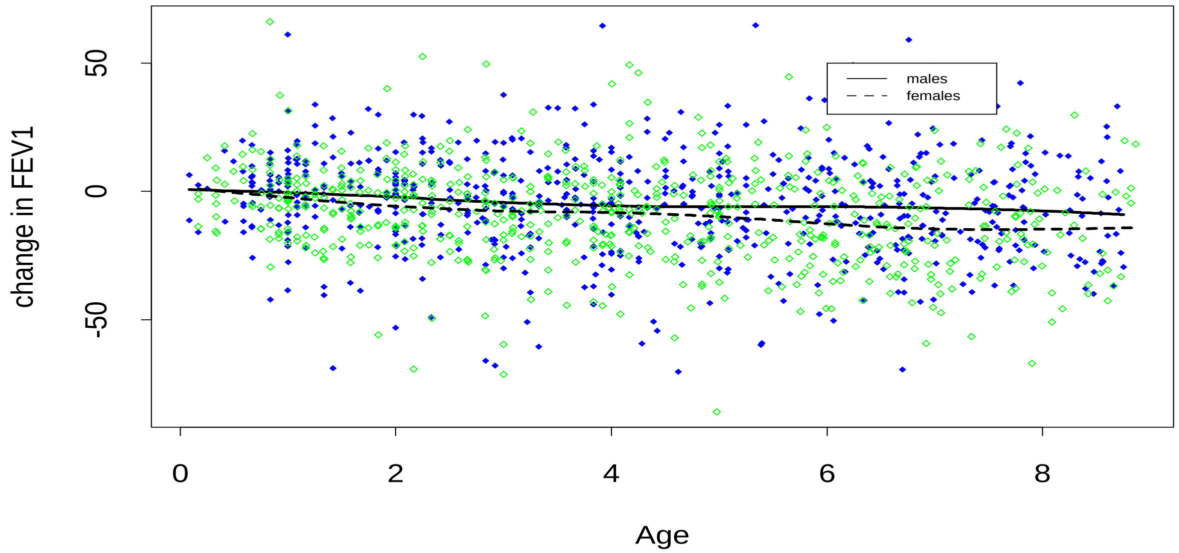
$$(Y_{ij} - Y_{i1}) = \beta_L (X_{ij} - X_{i1}) + (\epsilon_{ij} - \epsilon_{i1}).$$

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

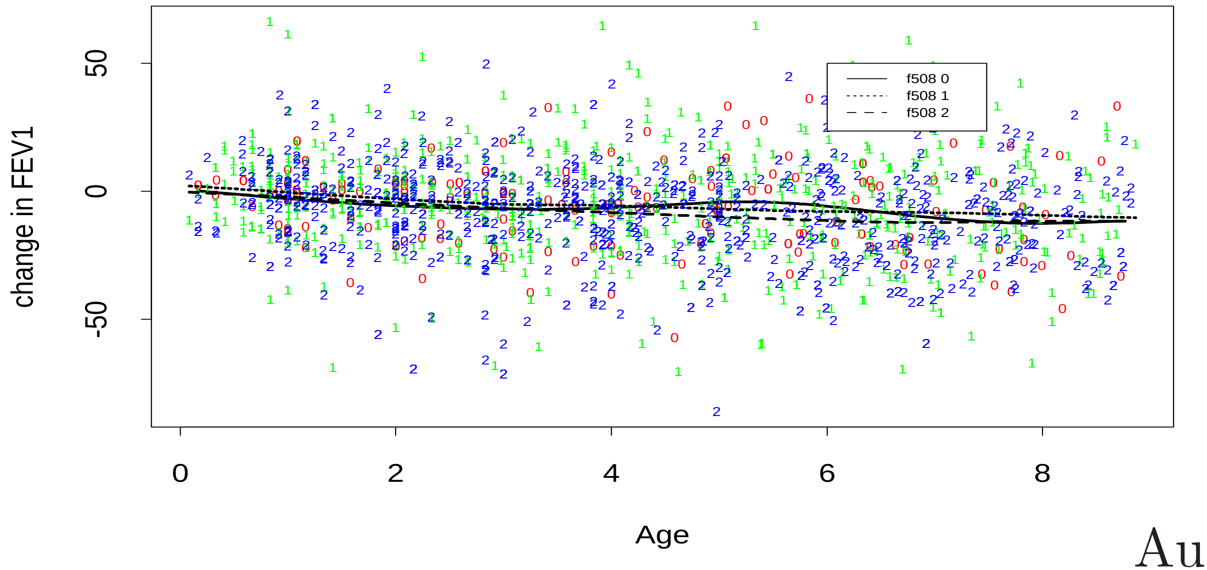
- **EDA**: plot $Y_{ij} - Y_{i1}$ versus $X_{ij} - X_{i1}$.


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

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?
- **Models for analysis**?

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.

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

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