Chapter 6

Statistical Analysis with Incomplete Grouped Data

6.1 Missing Data in Longitudinal Studies

A major challenge for the analysis of grouped/longitudinal data is missing data

- ▷ Even though studies are designed to collect data on every subject at a set of prespecified follow-up times
- ▷ Subjects often miss some of their planned measurements for a variety of reasons

6.1 Missing Data in Longitudinal Studies (cont'd)

- Implications of missingness:
  - ▷ we collect less data than originally planned ⇒ *loss of efficiency*
  - ▷ not all subjects have the same number of measurements ⇒ *unbalanced datasets*
  - ▷ missingness may depend on outcome ⇒ *potential bias*

- For the handling of missing data, we introduce the missing data indicator

$$r_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

6.1 Missing Data in Longitudinal Studies (cont'd)

- We obtain a partition of the complete response vector y_i
  - ▷ observed data y_i^o, containing those y_{ij} for which $r_{ij} = 1$
  - ▷ missing data y_i^m, containing those y_{ij} for which $r_{ij} = 0$

- We can have different patterns of missing data...

6.1 Missing Data in Longitudinal Studies (cont'd)

Subject	Visits				
	1	2	3	4	5
1	x	x	x	x	x
2	x	x	x	?	?
3	?	x	x	x	x
4	?	x	?	x	?

- ▷ Subject 1: Completer
- ▷ Subject 2: dropout
- ▷ Subject 3: late entry
- ▷ Subject 4: intermittent

6.1 Missing Data in Longitudinal Studies (cont'd)

- When the focus is only on dropout the notation can be simplified
  - ▷ Discrete dropout time: $r_i^d = 1 + \sum_{j=1}^{n_i} r_{ij}$ (ordinal variable)
  - ▷ Continuous dropout time: T_i^* denotes the time to dropout

- Focusing on dropout only is justifiable because often intermittent missing data can be considered MAR (definition of MAR follows...)

6.2 Missing Data Mechanisms

- To describe the probabilistic relation between the measurement and missingness processes Rubin (1976, Biometrika) has introduced three mechanisms

- *Missing Completely At Random (MCAR)*: The probability that responses are missing is unrelated to both y_i^o and y_i^m

$$p(r_i \mid y_i^o, y_i^m) = p(r_i)$$

- Examples
  - ▷ subjects go out of the study after providing a pre-determined number of measurements
  - ▷ laboratory measurements are lost due to equipment malfunction

6.2 Missing Data Mechanisms (cont'd)

- Features of MCAR:
  - ▷ The observed data y_i^o **can** be considered a random sample of the complete data y_i
  - ▷ We can use any statistical procedure that is valid for complete data
    - * sample averages per time point
    - * linear regression, ignoring the correlation (consistent, but not efficient)
    - * t-test at the last time point
    - * ...

6.2 Missing Data Mechanisms (cont'd)

- *Missing At Random (MAR)*: The probability that responses are missing is related to y_i^o, but is unrelated to y_i^m

$$p(r_i \mid y_i^o, y_i^m) = p(r_i \mid y_i^o)$$

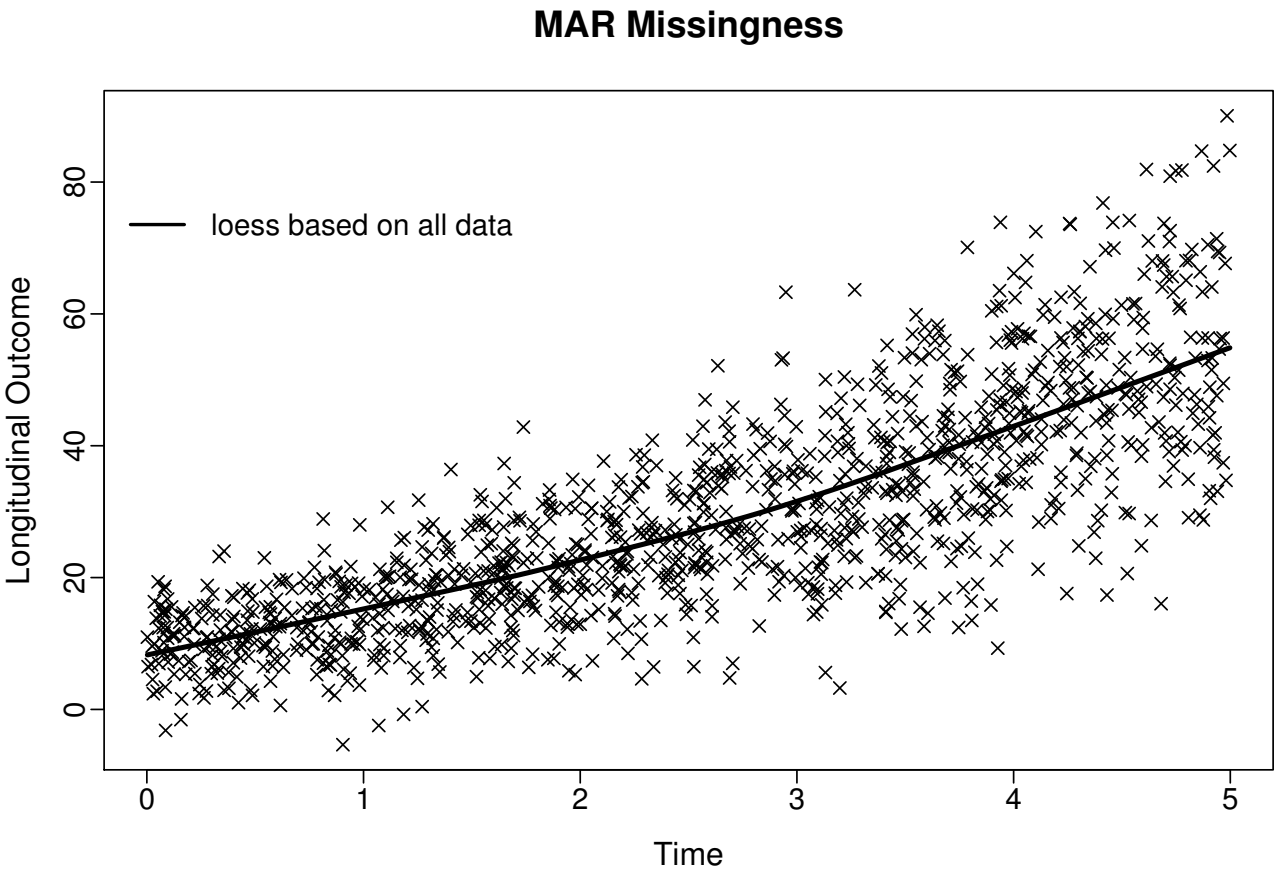
- Examples
  - ▷ study protocol requires patients whose response value exceeds a threshold to be removed from the study
  - ▷ physicians give rescue medication to patients who do not respond to treatment

6.2 Missing Data Mechanisms (cont'd)

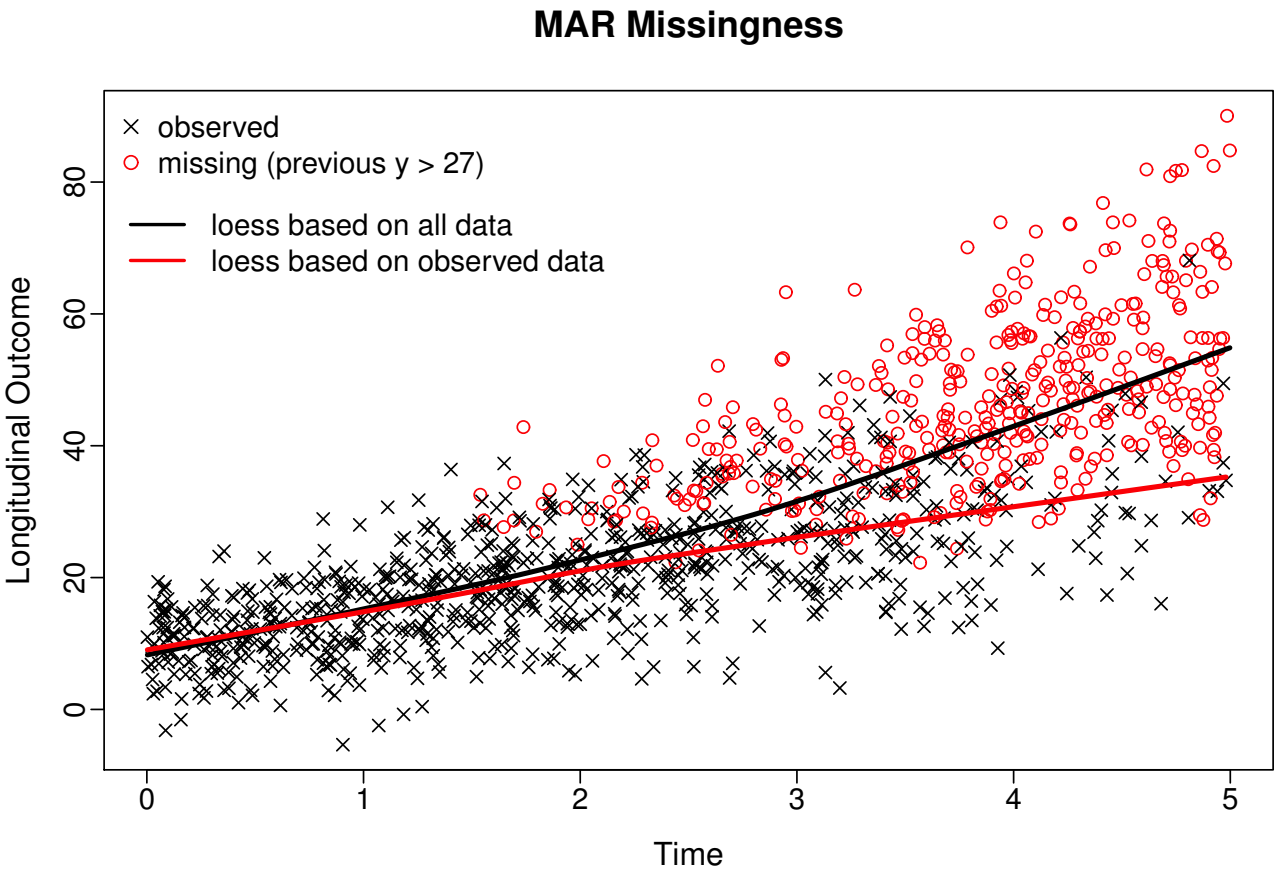
- Features of MAR:
  - ▷ The observed data **cannot** be considered a random sample from the target population
  - ▷ Not all statistical procedures provide valid results

Not valid under MAR	Valid under MAR
sample marginal evolutions	sample subject-specific evolutions
methods based on moments, such as GEE	likelihood based inference
multivariate models with misspecified correlation structure	multivariate models with correctly specified correlation structure
marginal residuals	subject-specific residuals

6.2 Missing Data Mechanisms (cont'd)



6.2 Missing Data Mechanisms (cont'd)



6.2 Missing Data Mechanisms (cont'd)

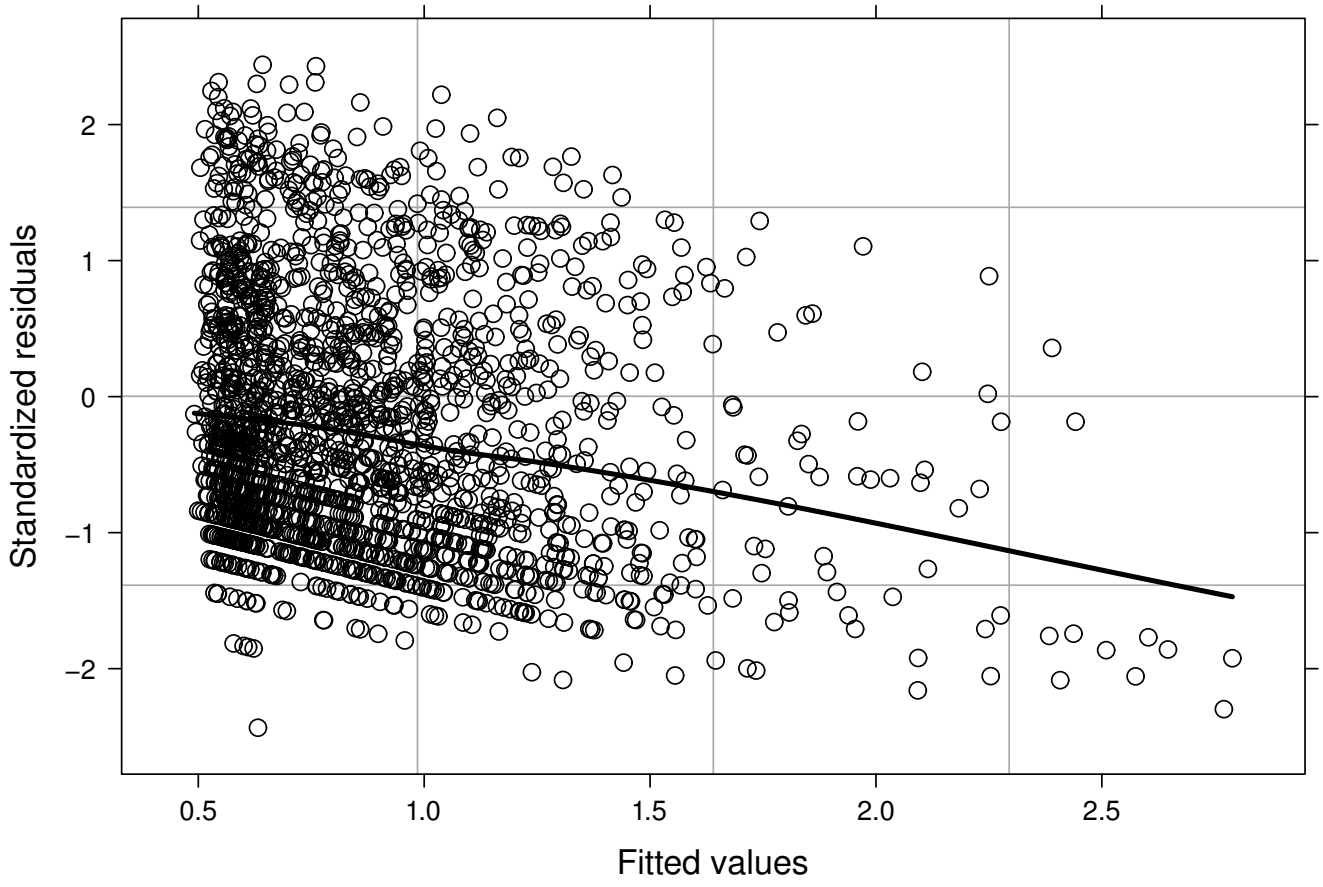
- To illustrate the important implications of incomplete data, let's return to the residuals plots we have seen in Chapter 2 (pp.131–135)

- We had fitted the following model to the PBC dataset

$$\begin{cases} \log(\texttt{serBilir}_{ij}) = \beta_0 + \beta_1 \texttt{Time}_{ij} + \beta_2 \texttt{Female}_i + \beta_3 \texttt{Age}_i + \\ \qquad\qquad \beta_4 \{\texttt{D-penicil}_i \times \texttt{Time}_{ij}\} + \beta_5 \{\texttt{Female}_i \times \texttt{Time}_{ij}\} + \varepsilon_{ij} \\ \varepsilon_i \sim \mathcal{N}(0, V_i) \qquad\qquad V_i \text{ has a continuous AR1 structure} \end{cases}$$

and the scatterplot of the standardized residuals versus fitted values was

6.2 Missing Data Mechanisms (cont'd)



6.2 Missing Data Mechanisms (cont'd)

- We see a clear systematic trend

- What's the problem?
  - ▷ is this really a model misspecification, or
  - ▷ is it an artefact of missing data?

- Why we say that:
  - ▷ patients with high serum bilirubin levels have higher chance of dropping out
  - ▷ the model will account for that and give as average longitudinal evolution the average of patients who did not drop out (i.e., observed evolutions), and the patients who did drop out (i.e., unobserved evolutions)

6.2 Missing Data Mechanisms (cont'd)

- However, the residuals are calculated based on the observed data alone

- Hence, even if the model is correct, we could still see systematic trends because of dropout

With MAR incomplete data standard residuals plots may show misleading systematic trends

6.2 Missing Data Mechanisms (cont'd)

- *Missing Not At Random (MNAR)*: The probability that responses are missing is related to y_i^m, and possibly also to y_i^o

$$p(r_i \mid y_i^m) \quad \text{or} \quad p(r_i \mid y_i^o, y_i^m)$$

- Examples
  - ▷ in studies on drug addicts, people who return to drugs are less likely than others to report their status
  - ▷ in longitudinal studies for quality-of-life, patients may fail to complete the questionnaire at occasions when their quality-of-life is compromised

6.2 Missing Data Mechanisms (cont'd)

- Features of MNAR
  - ▷ The observed data **cannot** be considered a random sample from the target population
  - ▷ Only procedures that explicitly model the joint distribution $\{y_i^o, y_i^m, r_i\}$ provide valid inferences $\Rightarrow$ **analyses which are valid under MAR will not be valid under MNAR**

6.2 Missing Data Mechanisms (cont'd)

We cannot tell from the data at hand whether the missing data mechanism is MAR or MNAR

Note: We can distinguish between MCAR and MAR

6.2 Missing Data Mechanisms (cont'd)

- *Missing Covariate Depended*: The probability that responses are missing is related to covariates x

$$p(r_i \mid x_i, y_i^o, y_i^m) = p(r_i \mid x_i)$$

- Examples
  - ▷ in a study on hypertensive patients, overweight patients are inclined not to have their blood pressure measured, and BMI is related with blood pressure

6.2 Missing Data Mechanisms (cont'd)

- Features of Missing Covariate Depended
  - ▷ If we do not include the covariates that drive the missingness process in the regression model for the longitudinal outcome Y, and these covariates are associated with Y, then we obtain an MNAR mechanism

6.3 Analysis with Incomplete Data

- We have seen what the implications of missingness are and how they complicate matters

- To this end, several approaches have been proposed to account for missing data
  - ▷ **depending on the missing data mechanism, not all of them provide valid results!**

6.3 Analysis with Incomplete Data (cont'd)

- **Complete Cases Analysis**
  - ▷ **General idea:** Restrict analyses to only those subjects for which all measurements are observed
  - ▷ **Advantages:**
    * very simple to implement
    * standard software can be used

    Disadvantages:
    * substantial loss of information
    * valid inferences only when missingness is completely unrelated to the outcome (i.e., MCAR)

6.3 Analysis with Incomplete Data (cont'd)

- **Last Observation Carried Forward (LOCF)**
  - ▷ **General idea:** Any missing value is replaced by the last observed value
  - ▷ **Advantages:**
    * very simple to implement
    * standard software can be used

    Disadvantages:
    * extremely strong assumption that a subject's measurement stays at the same level as soon as he/she is not observed
    * even if the mechanism is MCAR, LOCF may not provide valid results
    * overestimates precision

6.3 Analysis with Incomplete Data (cont'd)

- **Unconditional Mean Imputation**
  - ▷ **General idea:** Each missing outcome y_{ij}^m is replaced by the average of the observed measurements at the j-th occasion
  - ▷ **Advantages:**
    * very simple to implement
    * standard software can be used

    Disadvantages:
    * can only be implemented with balanced designs
    * it provides valid results only under MCAR
    * overestimates precision

6.3 Analysis with Incomplete Data (cont'd)

- **Conditional Mean Imputation**
  - ▷ **General idea:** The vector y_i^m of missing observations for the i-th subject is replaced by its prediction, conditional on the vector y_i^o of observed observations for that subject
    * we specify a model for y_i^m conditional on y_i^o and parameters θ – often this model will result from a full specification of the marginal model $y_i = (y_i^o, y_i^m)$
    * we fit the model to the completers and obtain estimates $\widehat{\theta}$ for the parameters
    * based on this fitted model we can calculate predictions for the missing observations, i.e.,

$$\widehat{y_i^m} = E(y_i^m \mid y_i^o, \widehat{\theta})$$

6.3 Analysis with Incomplete Data (cont'd)

- **Conditional Mean Imputation**
  - ▷ **Advantages:**
    * less strict assumptions that the previously mentioned approaches

    Disadvantages:
    * requires programming for its implementation
    * overestimates precision

6.3 Analysis with Incomplete Data (cont'd)

- **Multiple Imputation**
  - ▷ A common issue in all aforementioned imputation techniques was the *overestimation of precision*
    $\Rightarrow$ no correction was made for the uncertainty introduced from imputing the missing observations
  - ▷ **General idea:** To propagate this uncertainty we impute not only once but *multiple* times from the conditional distribution $p(y_i^m \mid y_i^o, \widehat{\theta})$
    * M completed datasets are formed
    * we perform the same analysis in each
    * we pool the estimated parameters using Rubin's formulas

6.3 Analysis with Incomplete Data (cont'd)

- **Multiple Imputation**
  - ▷ **Advantages:**
    * correctly propagates uncertainty due to incomplete data
    * valid under MAR
    * allows for different types of analysis (e.g., concentrate at a specific time point – cross-sectional analysis)

  Disadvantages:
    * not available for grouped/clustered data in all software

6.3 Analysis with Incomplete Data (cont'd)

- **Full Specification of the Outcome Distribution**
  - ▷ **General idea:** Use a model for the joint distribution of the responses – this includes the models we have seen in Chapter 2, 3, & 5 (but not the GEE approach of Chapter 4)
  - ▷ **Advantages:**
    * no requirement to impute data
    * available in all standard software
    * valid results under MCAR and MAR

    Disadvantages:
    * not valid results under MNAR

6.3 Analysis with Incomplete Data (cont'd)

- **Missing Not At Random Models**
  - ▷ **General idea:** When the missing data mechanism is MNAR, we need to define a model for the joint distribution of the longitudinal outcome $\{y_i^o, y_i^m\}$ and the missingness outcome r_i

    * Three model families have been proposed
      - *selection models*
      - *pattern mixture models*
      - *shared parameter models*

6.3 Analysis with Incomplete Data (cont'd)

- **Missing Not At Random Models**

  * Selection models use the decomposition

$$p(y_i^o, y_i^m, r_i) = p(y_i^o, y_i^m)\, p(r_i \mid y_i^o, y_i^m)$$

  * These models postulate that the probability of dropping out is directly related on the missing longitudinal outcomes

6.3 Analysis with Incomplete Data (cont'd)

- **Missing Not At Random Models**

  * Pattern mixture models use the decomposition

$$p(y_i^o, y_i^m, r_i) = p(y_i^o, y_i^m \mid r_i)\, p(r_i)$$

  * These models postulate that we have a different specification of the longitudinal model per dropout pattern (e.g., completers show different average evolutions than subjects who dropout earlier on)

6.3 Analysis with Incomplete Data (cont'd)

- **Missing Not At Random Models**

  * Shared parameter models use the decomposition

$$p(y_i^o, y_i^m, r_i) = \int p(y_i^o, y_i^m \mid b_i)\, p(r_i \mid b_i)\, p(b_i)\, db_i$$

$$= \int p(y_i^o, \mid b_i)\, p(r_i \mid b_i)\, p(b_i)\, db_i$$

  * These models postulate that the characteristics of the longitudinal profile of a subject (described by the random effects) dictate the chance of dropping out

6.3 Analysis with Incomplete Data (cont'd)

- **Missing Not At Random Models**
  - ▷ **Advantages:**
    * no requirement to impute data
    * provide valid results under MNAR

  Disadvantages:
    * only some of them available in software
    * difficult to fit
    * require sensitivity analysis

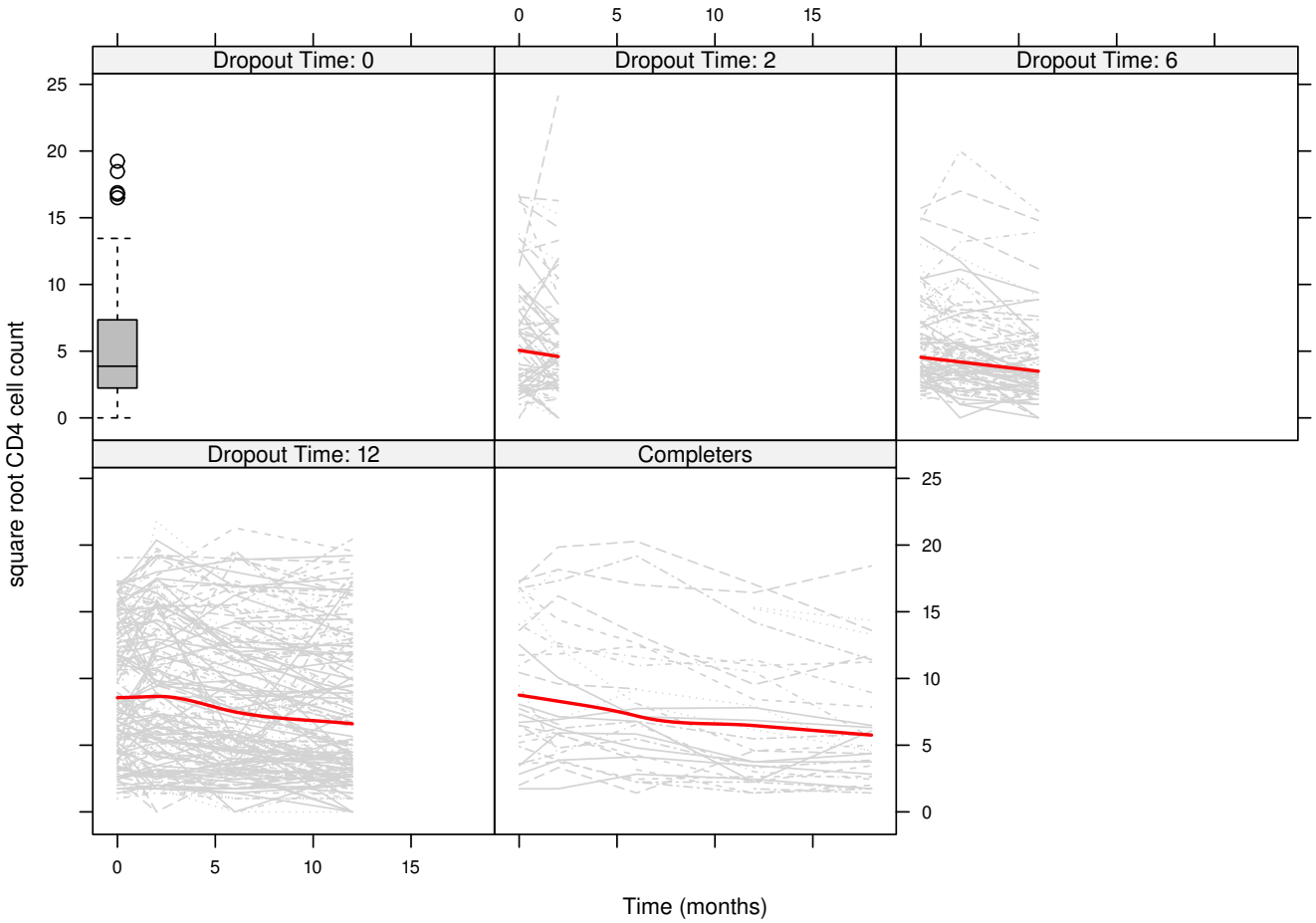
6.3 Analysis with Incomplete Data (cont'd)

- Example: In the AIDS data set we have a considerable amount of missing data

	Missing Data per Month				
	0	2	6	12	18
Freq.	0	99	157	241	433
%	0.0	21.2	33.6	51.6	92.7

- The sample evolutions of the square root CD4 cell counts per dropout pattern have the form

6.3 Analysis with Incomplete Data (cont'd)



6.3 Analysis with Incomplete Data (cont'd)

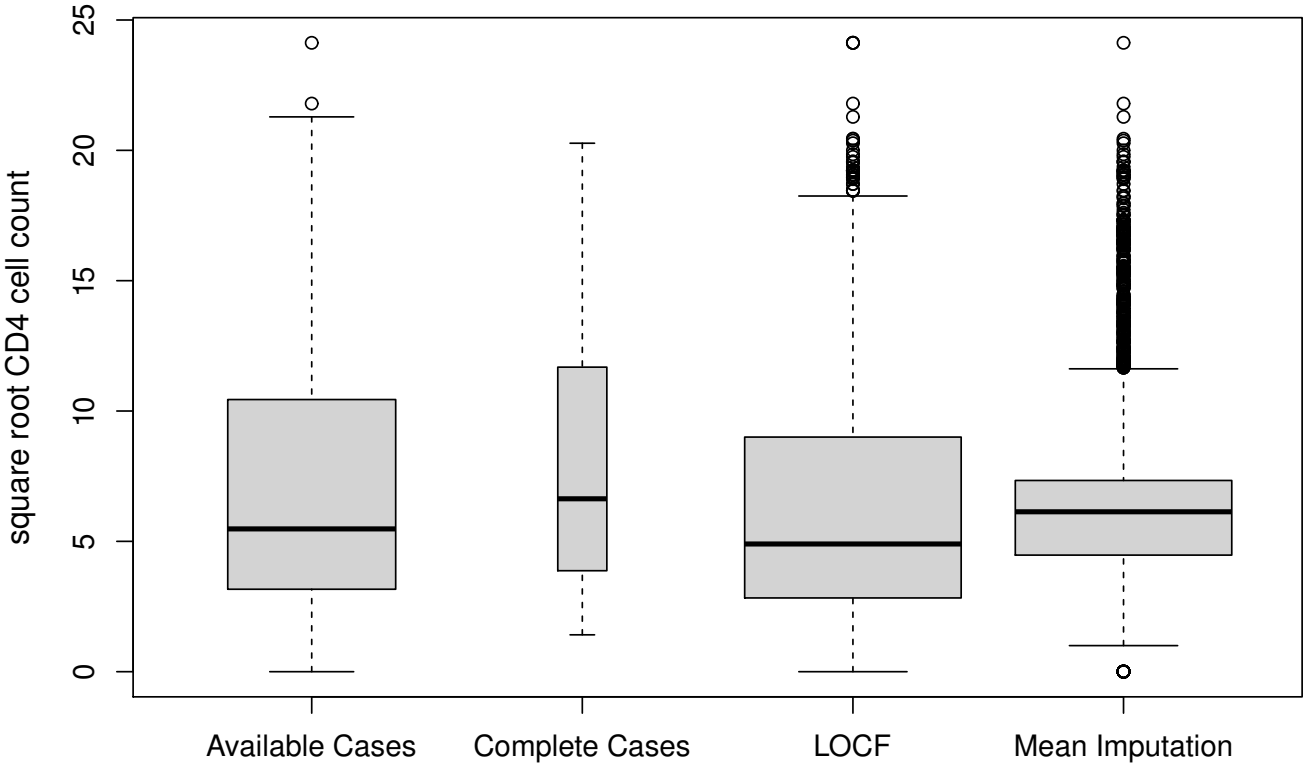
- We are interested in a mixed model with the following specification
  - ▷ fixed effects:
    - * main effects: time, AZT and previous opportunistic infection
    - * interaction effects: time with AZT, and time with previous opportunistic infection
  - ▷ random effects: random intercepts & random slopes

- We will compare the MAR analysis (using a linear mixed model and all available cases) with
  - ▷ complete cases analysis
  - ▷ last observation carried forward analysis
  - ▷ Mean Imputation analysis

6.3 Analysis with Incomplete Data (cont'd)

- The following boxplots illustrate the distribution of square root CD4 cell counts under the different strategies

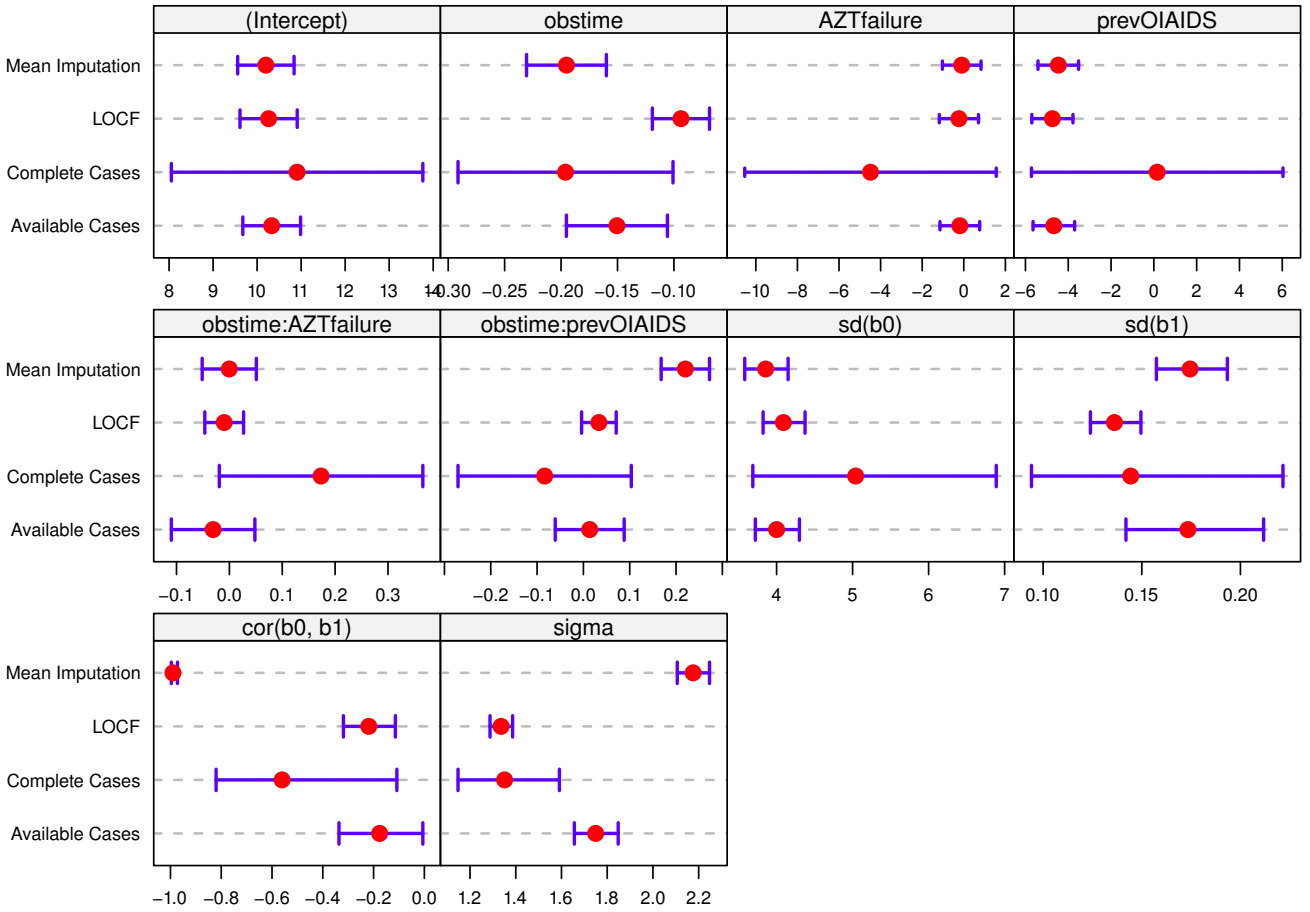
6.3 Analysis with Incomplete Data (cont'd)



6.3 Analysis with Incomplete Data (cont'd)

- The following figure illustrates the estimated coefficients and the corresponding 95% confidence intervals from the four mixed models fitted to the different versions of the square root CD4 cell counts variable

6.3 Analysis with Incomplete Data (cont'd)



6.3 Analysis with Incomplete Data (cont'd)

- We observe considerable differences between the different approaches with respect to both
  - ▷ parameter estimates
  - ▷ standard errors (width of the confidence intervals)

The manner one decides to handle incomplete data can have a profound effect in the derived results

6.4 Summary

- It is now universally recognized (i.e., officially also by FDA) that the default type of statistical analysis should provide valid results under MAR

- Hence, whenever we have missing data in the outcome, it is advisable to employ a full-likelihood approach based on the models we have seen for continuous and categorical responses
  $\Rightarrow$ **No need for (multiple) imputation**

- *However,* to be protected we need an appropriate specification of the joint distribution of the data

6.4 Summary (cont'd)

- This encompasses both the mean and the covariance/correlation structure
  $\Rightarrow$ **do not favor simpler covariance matrices if the p-value is just non-significant**

- When we also have missing data in the covariates a Multiple Imputation approach should be employed

- In general, when dealing with incomplete data it is advisable to perform a **Sensitivity Analysis**

  ▷ check how results change under logical alterations of your model

6.5 Review of Key Points

- Missing data pose an important complication in the analysis of clustered/grouped data
  - ▷ loss of efficiency
  - ▷ potential bias

- Need to carefully consider the reasons why data are missing
  - ▷ MCAR $\Rightarrow$ missingness not related to the outcome
  - ▷ MAR $\Rightarrow$ missingness related to the *observed* part of the outcome
  - ▷ MNAR $\Rightarrow$ missingness related to the *unobserved* part of the outcome

6.5 Review of Key Points (cont'd)

- Standard analysis should be one that provides valid results under (at least) MAR
  - ▷ full & flexible specification of the distribution of the data
  - ▷ weighted GEE
  - ▷ sensitivity analysis

- MNAR setting $\Rightarrow$ difficult to handle in practice
  - ▷ in some cases, including important covariates may alleviate the problem
  - ▷ missing covariate depended

Chapter 7

Closing

7.1 Concluding Remarks

- **Features of cluster/grouped data**
  - ▷ measurements in the same cluster are correlated
  - ▷ distinction of between units and within units effects
  - ▷ often some measurements are missing for various reasons

Statistical techniques that ignore these features may produce spurious results

7.1 Concluding Remarks (cont'd)

- **Two major modeling frameworks**
  - ▷ marginal models
  - ▷ mixed effects models

- **Continuous vs discrete data**
  - ▷ continuous/normal data $\Rightarrow$ mixed models imply a specific marginal model
  - ▷ discrete data $\Rightarrow$ more substantial differences between the two frameworks

7.1 Concluding Remarks (cont'd)

- **Missing data**
  - ▷ careful consideration of the missing data mechanism (i.e., reasons why the data are missing)
  - ▷ default should be an MAR analysis + sensitivity analysis

- **What we did not cover?**
  - ▷ multivariate Poisson & multivariate ordinal data
  - ▷ nonlinear models for multivariate data
  - ▷ transition models
  - ▷ alternating logistic regression
  - ▷ weighted GEE & doubly robust methods

The End!