Chapter 5

Statistical Analysis with Incomplete Grouped Data

5.1 Missing Data in Longitudinal Studies

- A major challenge for the analysis of longitudinal data is the problem of missing data
  - ▷ studies are designed to collect data on every subject at a set of prespecified follow-up times
  - ▷ often subjects miss some of their planned measurements for a variety of reasons

- We can have different patterns of missing data

5.1 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

5.1 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

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

5.1 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 $r_{ij} = 0$

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

5.2 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

5.2 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
    - * $\ldots$

5.2 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

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

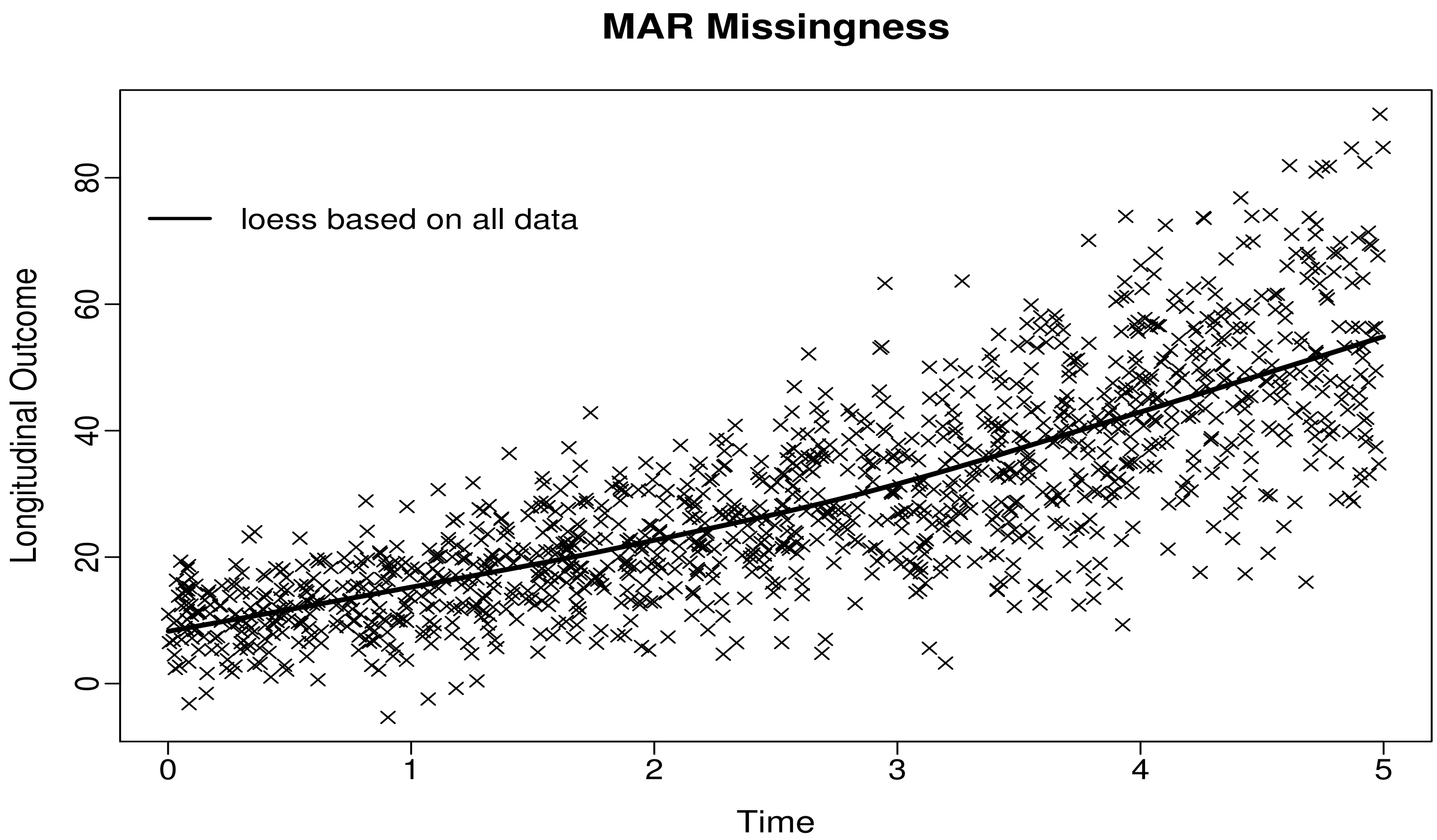
- Examples
  - ▷ study protocol requires patients whose response value exceeds a threshold to be removed from the study
  - ▷ physicians give rescue medication to patients who do not respond to treatment

5.2 Missing Data Mechanisms (cont'd)

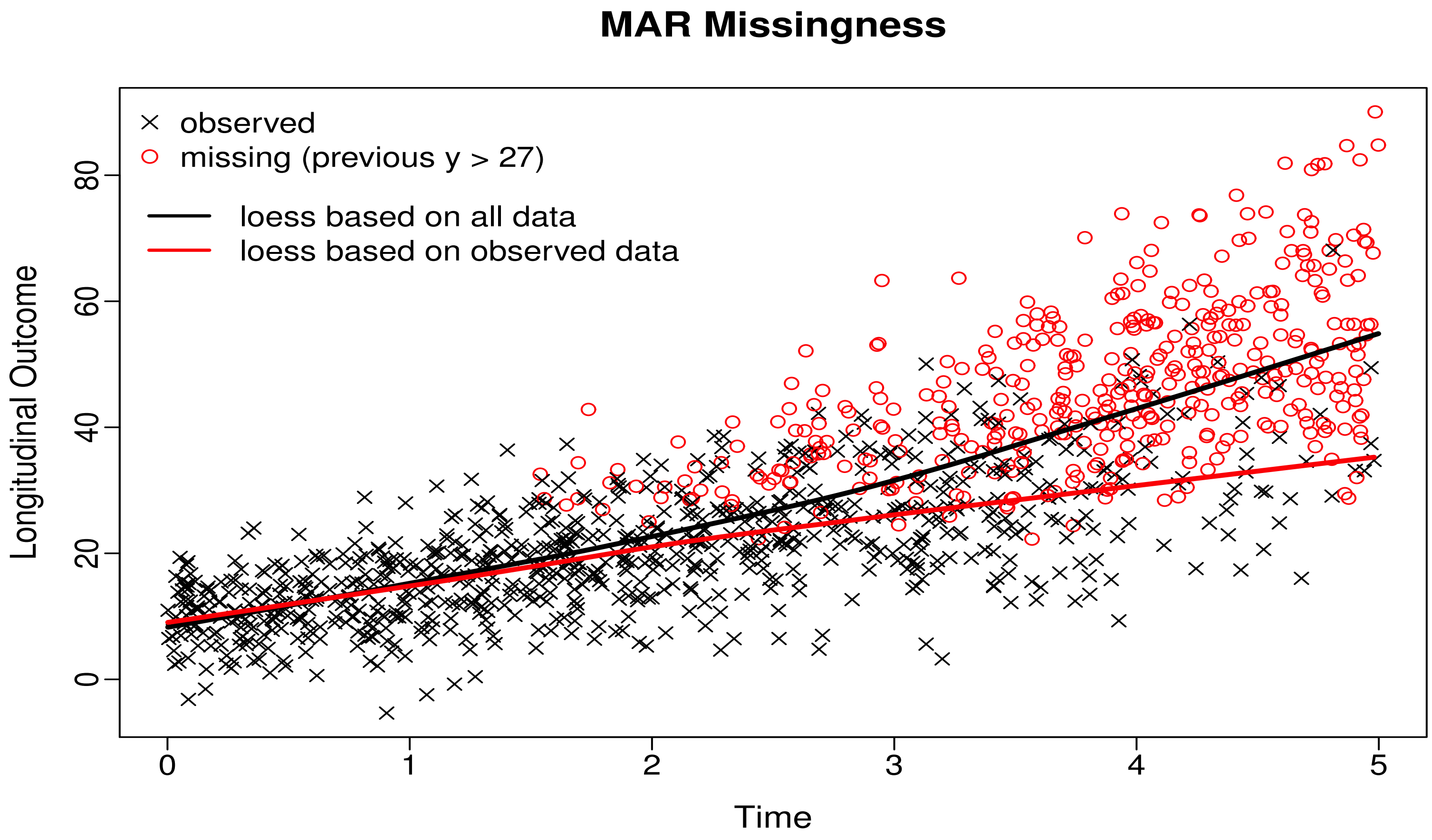
- Features of MAR:
  - ▷ The observed data cannot be considered a random sample from the target population
  - ▷ Not all statistical procedures provide valid results

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

5.2 Missing Data Mechanisms (cont'd)



5.2 Missing Data Mechanisms (cont'd)



5.2 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

5.2 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 ⇒ **analyses which are valid under MAR will not be valid under MNAR**

5.2 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

5.3 Review of Key Points

- ***