Introduction to Missing Data

Missing Data in Longitudinal Studies

When do missing data occur?

Intended measurements are not taken, or lost, or are not available.

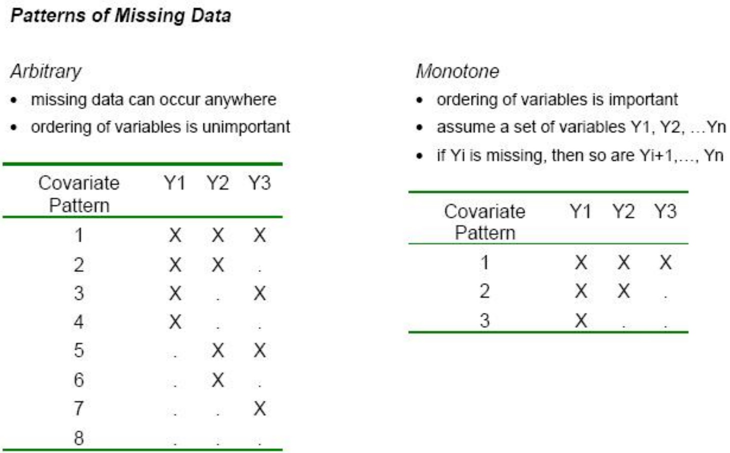
Patterns of missing data in longitudinal studies:

- Dropouts: Subjects stop participating in the study during the followup period and never come back. Reasons could be related to poor treatment outcomes, death, cure, loss of interest, moving away, etc.
- Intermittent Missing: Subjects have missing values in the middle of a study, e.g., miss an appointment, but come back later during the study.

Remarks:

It is important to understand the **reasons** for missing data, as valid statistical inference in the presence of missing data strongly depend on such assumptions.

Patterns of Missingness



Patterns of Missing Data

Arbitrary

- missing data can occur anywhere
- ordering of variables is unimportant

Covariate Pattern	Y1	Y2	Y3
1	X	X	X
2	X	X	.
3	X	.	X
4	X	.	.
5	.	X	X
6	.	X	.
7	.	.	X
8	.	.	.

Monotone

- ordering of variables is important
- assume a set of variables Y1, Y2, …Yn
- if Yi is missing, then so are Yi+1,…, Yn

Covariate Pattern	Y1	Y2	Y3
1	X	X	X
2	X	X	.
3	X	.	.

Why do we care?

- Missing data may cause biased estimates if handled incorrectly
- The presence of bias and the degree of bias depend on the nature of the missing data process and the analysis methods
- Analysis of missing data in longitudinal studies requires specification of the missing data mechanism and often requires joint modeling of the outcome (Y) and the missing data indicator R.

Why 571? Because missingness frequently arises in longitudinal data. It's hard to keep everyone in the study or get them to show up at the right times.

Example 1: Hedeker and Gibbons, 1997

- Randomized psychiatric trial
- 312 patients received drug therapies for schizophrenia; 101 patients received a placebo (3:1 randomization)
- Measurements at weeks 0, 1, 3, 6
- Missing data primarily due to dropout
- Outcome: severity of illness (1=normal, . . . , 7=extremely ill); treated as continuous

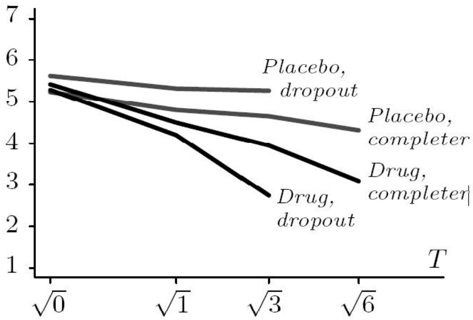
Trial sample size:

		Time		
Group	0	1	3	6
Placebo ($n = 108$)	107	105	87	70
Drug ($n = 329$)	327	321	287	265

Note: The drug group combines three treatments.

Dropout Rates: Placebo: 35% & Drug: 20%.

Average response versus square root of time (in weeks):



- In the treatment group, the subjects who dropped out had lower scores than the completers.
- In the control group, the subjects who dropped out had higher scores than the completers.
- A completer-only (complete case) analysis would severely understate the treatment effect.

References

General references on missing data:

- Little and Rubin (2002) Statistical Analysis with Missing Data, 2nd edition. Wiley.
- Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data. Chapman & Hall.

References on missing data in longitudinal studies:

- Diggle and Kenward (1994), Informative dropouts in longitudinal Studies, JRSSC, 43, 47-73.
- Little, R.J. (1995) Modeling the dropout mechanism in repeated-measures studies. JASA
- Diggle, et al. (2004) Analysis of Longitudinal Data, 2nd Edition. Cambridge.
- Verbeke, G. and Molenberghs, G. (2000) Linear Mixed Models for Longitudinal Data. Springer.

Missing Data Mechanisms

Missing Mechanisms (Little and Rubin, 2002):

- Missing Completely At Random (MCAR): Missingness does not depend on outcomes and covariates.
- Missing At Random (MAR): Missingness only depends on observed outcomes and covariates.
- Nonignorable (Informative) Missing(NMAR): Missingness depends on unobserved outcomes or unobserved covariates.

Missing Data Mechanisms (2)

- $\boldsymbol{Y}_{obs}$: Measurements observed.
- $\boldsymbol{Y}_{mis}$: Measurements that should be available but are missing.
- $\mathbf{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$: Hypothetical complete data.
- $\mathbf{R}$: Indicator of whether Y is observed or not, or the time when Y is missing.
- Assume the covariates $\mathbf{X}$ are fully observed.

Example: Assume three time points ($n = 3$) and suppressing subscript i, denote the three outcomes by (Y_1, Y_2, Y_3)

Pattern	Y_1	Y_2	Y_3
P1	X	X	X
P2	X	X	
P3	X		

Missing Data Mechanisms (3)

Notation:

Subjects having Pattern 1 (complete cases): $Y_{obs} = (Y_1, Y_2, Y_3)$, no Y_{mis}, $R = (1, 1, 1)$ or $R = 4$ (subjects have missing data at time 4).

Subjects having Pattern 2: $Y_{obs} = (Y_1, Y_2)$ and $Y_{mis} = Y_3$, $R = (1, 1, 0)$ or $R = 3$ (subjects have missing data starting from time 3.

Subjects having Pattern 3: $Y_{obs} = Y_1$ and $Y_{mis} = (Y_{1,} Y_2)$, $R = (1, 0, 0)$ or $R = 2$ (subjects have missing data starting from time 2.

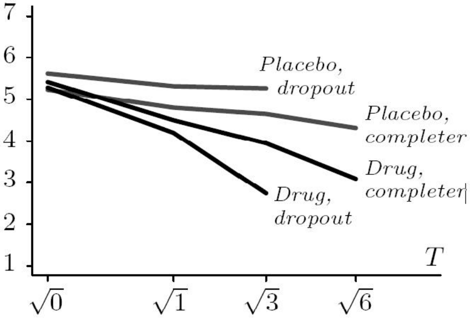
Missing Data Mechanisms (4)

- Missing Completely At Random (MCAR): if $\mathbf{R}$ is independent of both $\boldsymbol{Y}_{obs}$ and $\boldsymbol{Y}_{mis}$.
  - Patients miss a scheduled visit because of bad weather or car out of service.

- Missing At Random (MAR): if $\mathbf{R}$ is independent of $\boldsymbol{Y}_{mis}$, but dependent on $\boldsymbol{Y}_{obs}$, i.e., the probability of missingness only depends on Y_{obs}, but not Y_{mis}.
  - Older people may have a higher chance of dropping out of a study. (Suppose age is observable.)

- Nonignorable (Informative) Missing(NMAR): if $\mathbf{R}$ is dependent on $\boldsymbol{Y}_{mis}$, i.e., the probability of missingness depends on both $Y_{mis}$and Y_{obs}.
  - Subjects drop out because they have poor treatment outcomes or they die.

Example: Psychiatric Trial Revisited:



- Dropout is not MCAR, because it operates differently in the treatment and control groups.
- Dropout is not merely related to covariates, because completers and dropouts follow different (pre-dropout) trajectories.
- Dropout could be MAR or nonignorable; it's impossible to tell from the data.

Naive Approaches to Missing Data

"Easy" approaches to dealing with missingness (just getting the software to run!)

1. Mean imputation
2. Last observation carried forward for dropout
3. Complete case analysis

Mean Imputation

Y_{ij}: response for subject i at the jth time point

$$R_{ij} = \begin{cases} 1, & observed \\ 0, & missing \end{cases}$$

If Y_{ij} is missing, replace it by

- the mean response for subject i

$$Y_{i\cdot} = \frac{\sum_j R_{ij} Y_{ij}}{\sum_j R_{ij}}$$

- the mean response for occasion j

$$Y_{\cdot j} = \frac{\sum_i R_{ij} Y_{ij}}{\sum_i R_{ij}}$$

- Some predicted value

Note: These methods may (not always!) seriously distort estimates and measures of uncertainty.

Last observation carried forward for dropout

Idea: If ith subject drops out after occasion j, replace (impute) $Y_{i,j+1}$, $Y_{i,j+2}$,.... by Y_{ij}.

Remarks:

- (Was?) Routinely used in the pharmaceutical industry for randomized trials.
- Last observation carried forward tends to understate differences in estimated time-trends between treatment and control groups (tend to be attenuated).
- Inference is not necessarily conservative, because standard errors are biased downward as well.
- It performs especially poorly for outcomes that have high variation within a subject.
- Not recommended as a general method.

Last observation carried forward for dropout

- Single-imputation strategies designed to precisely predict the missing values tend to distort estimates of population quantities.

- The goal of the missing-data analysis is usually to draw accurate inferences about population quantities (e.g. mean change over time), not to accurately predict missing values.

- Using imputation methods, the best way to achieve this goal is to preserve all aspects of the data distribution (means, trends, within- and between-subject variation, etc.)

- Ad hoc imputation methods (like the above two methods) inevitably preserve some aspects but distort others.

Complete case analysis

- Discard all incomplete data and only use the complete cases to make inference.
- Potential to introduce bias if the dropout process is related to the measurement process and obviously wasteful of data.
- Not so bad for experiments for which data are nearly balanced and the missingness is often MAR or MCAR.
- For population inference, it's nearly always better to analyze all the available data from all the subjects, no matter whether they completed the study or not...
  - less biased
  - more efficient

Intermittent missing vs dropout

Dropout:

$$\begin{array}{lcccccc} \mathbf{Y}_i : & Y_{i1} & Y_{i2} & Y_{i3} & Y_{i4} & Y_{i5} & Y_{i6} \\ \mathbf{R}_i : & 1 & 1 & 0 & 0 & 0 & 0 \end{array}$$

Intermittent missing:

$$\begin{array}{lcccccc} \mathbf{Y}_i : & Y_{i1} & Y_{i2} & Y_{i3} & Y_{i4} & Y_{i5} & Y_{i6} \\ \mathbf{R}_i : & 1 & 1 & 0 & 1 & 1 & 0 \end{array}$$

Note: Intermittent missing is often harder to handle

Inference for Intermittent missing values vs dropouts

- When intermittent missing values arise through a known censoring mechanism, e.g. missing when below a known threshold in a lab test, one could use the EM algorithm (Dempster et al., 1977; Laird, 1988; Hughes, 1999).

- In practice it is often not to far off to assume the reasons for intermittent missing data are known and the missingnesss is unrelated to the measurement process. Then any analysis accommodating unbalanced data is usually valid.

- In contrast to intermittent missing data, it is more difficult to assume dropouts are unrelated to the measurement process. Then failing to properly modeling the missing data mechanism is more likely to result in biased inference and misleading results.

Overview of Statistical Models for Modeling Dropouts in Longitudinal Data

Likelihood Function:

$$L(Y_{obs}, R) = \int L(Y_{obs}, Y_{mis}, R) dY_{mis}$$

Two classes of models:

1. Selection model (Diggle and Kenward, 1994):

Partition the likelihood as

$$L(Y_{obs}, Y_{mis}, R) = L(Y_{obs}, Y_{mis}) L(R | Y_{obs}, Y_{mis})$$

2. Pattern Mixture Model (Little, 1993, 1994):

Partition the likelihood as

$$L(Y_{obs}, Y_{mis}, R) = L(R) L(Y_{obs}, Y_{mis} | R)$$

Selection Models and Pattern Mixture Models

1. The selection model and the pattern mixture model use different ways to partition the likelihood.
2. The selection model models the marginal distribution of the outcome and models the conditinal distribution of the dropout on the outcome. The regression coefficients from the Y model hence have attractive *population* interpretation in practice.
3. The pattern mixture model models the marginal distribution of the dropout and the conditional distributions of the outcome on each dropout pattern. Hence the interpretation of regression coefficients is conditional on the dropout patterns and has less attractive interpretation in practice.
4. However, pattern mixture models help researchers better understand the missing data mechanisms and assumptions.