Missing Data

Workshop: Analysis of Longitudinal Data

12th Nov 2024

Jaroslaw Harezlak

Armando Teixeira-Pinto

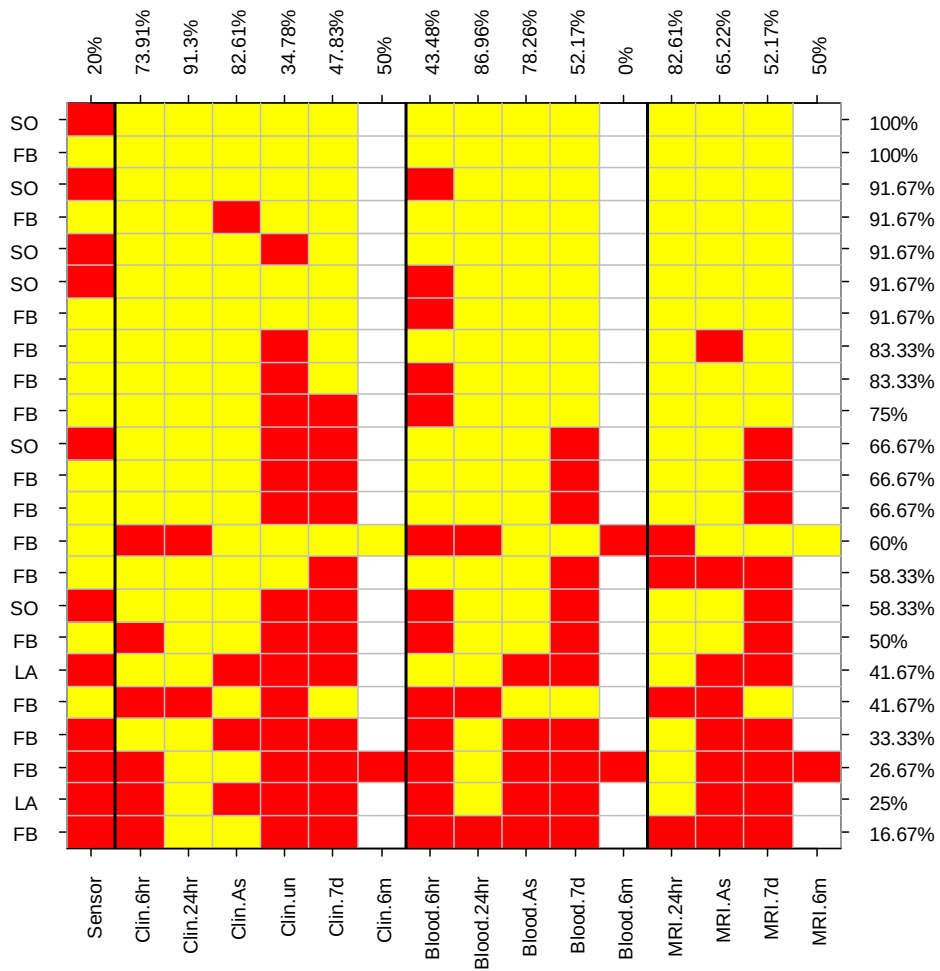
THE UNIVERSITY OF SYDNEY

Missing Data and Dropout

- Missing data arise in longitudinal studies whenever one or more of the sequences of measurements are incomplete, in the sense that some intended measurements are not obtained.

- In longitudinal studies missing data are the norm, NOT the exception

- Missing data:
  - Lead to loss of information
  - Can introduce bias and result in misleading inferences about change
  - Need to consider reasons for missingness


20%
73.91%
91.3%
82.61%
34.78%
47.83%
50%
43.48%
86.96%
78.26%
52.17%
0%
82.61%
65.22%
52.17%
50%
SO
FB
SO
FB
SO
SO
FB
FB
FB
FB
SO
FB
FB
FB
FB
FB
SO
FB
LA
FB
FB
FB
LA
FB
100%
100%
91.67%
91.67%
91.67%
91.67%
91.67%
83.33%
83.33%
75%
66.67%
66.67%
66.67%
60%
58.33%
58.33%
50%
41.67%
41.67%
33.33%
26.67%
25%
16.67%
Sensor
Clin.6hr
Clin.24hr
Clin.As
Clin.un
Clin.7d
Clin.6m
Blood.6hr
Blood.24hr
Blood.As
Blood.7d
Blood.6m
MRI.24hr
MRI.As
MRI.7d
MRI.6m

Missing data and dropout

Let $\mathbf{Y}$ denote the complete response vector which can be partitioned into two sub-vectors:

(i) $\mathbf{Y}^{(o)}$ the measurements observed
(ii) $\mathbf{Y}^{(m)}$ the measurements that are missing

If there were no missing data, we would have observed the complete response vector $\mathbf{Y}$.

Instead, we get to observe $\mathbf{Y}^{(o)}$.

Example

The main problem that arises with missing data is that the distribution of the observed data may not be the same as the distribution of the complete data.

Consider the following simple illustration:

- Suppose we intend to measure subjects at 6 months (Y_1) and 12 months (Y_2) post treatment.
- All subjects return for measurement at 6 months, but many do not return at 12 months.

If subjects fail to return for measurement at 12 months because they are not well (say, values of Y_2 are low), then the distribution of observed Y_2's will be positively skewed compared to the distribution of Y_2's in the population of interest.

In general, the situation may often be quite complex, with some missingness unrelated to either the observed or unobserved response, some related to the observed, some related to the unobserved, and some to both.

Missing data mechanisms

To obtain valid inferences from incomplete data the mechanism (probability model) producing the missing observations must be considered.

A hierarchy of three different types of missing data mechanisms can be distinguished:

1) Data are missing completely at random (**MCAR**) when the probability that an individual value will be missing is independent of $Y^{(o)}$ and $Y^{(m)}$.

2) Data are missing at random (**MAR**) when the probability that an individual value will be missing is independent of $Y^{(m)}$ (but may depend on $Y^{(o)}$).

3) Not Missing at Random (**NMAR**) (also called: non-ignorable) when the probability that an individual value will be missing depends on $Y^{(m)}$.

Note: Under assumptions 1) and 2), the missing data mechanism is often referred to as being `ignorable'.

Missing Completely at Random (MCAR)

MCAR: probability that responses are missing is unrelated to either the specific values that, in principle, should have been obtained (the missing responses) or the set of observed responses.

MCAR: probability responses are missing is independent of $Y^{(o)}$ and $Y^{(m)}$.

Missingness is simply the result of a chance mechanism that is unrelated to either observed or unobserved components of the outcome vector.

Consequently, observed data can by thought of as a random sample of the complete data.

Missingness depending on predictors

If missingness depends only on **X**, then technically it is MCAR. However, sometimes this is referred to as covariate dependent non-response.

Thus, in general, if non-response depends on covariates, **X**, it is harmless and the same as MCAR provided you always condition on the covariates (i.e., incorporate the covariate in the analysis). This type of missingness is only a problem if you do not condition on **X**.

Example: Consider the case where missingness depends on treatment group. Then the observed means in each treatment group are unbiased estimates of the population means.

However, the marginal response mean, averaged over the treatment groups, is not unbiased for the corresponding mean in the population (the latter, though, is usually not of subject-matter interest).

Stratification

Sometimes it may be necessary to introduce additional covariates, or stratifying variables, into the analysis to control for potential bias due to missingness.

Example: Suppose the response Y is a measure of health outcome, and X_1 is an indicator of treatment, and X_2 is an indicator of side-effects. Suppose missingness depends on side-effects.

If side-effects and outcome are uncorrelated, then there will be no bias.

If side-effects and outcome are correlated, then there will be bias unless you stratify the analysis on both treatment and side-effects (analogous to confounding).

Features of MCAR

The means, variances and covariances are preserved

If

$$E(Y_i) = X_i \beta \qquad \text{with complete data } Y_i - n \times 1,\ X_i - n \times p$$

then

$$E(Y_i) = I_i X_i \beta \quad = X_i \beta \qquad Y_i - n_i \times 1,\ X_i - n_i \times p$$

and

$$Cov(Y_i) = I_i \Sigma_i I'_i = \Sigma_i$$

where I_i is identity matrix with rows corresponding to missing values removed.

MCAR

Can use ML/REML estimators for β

In general, one can use GLS estimator with any "working" covariance matrix assumption

If working assumption on the covariance matrix made, need to use "empirical" ("sandwich") variance estimator for Var($\hat{\beta}$)

Any method valid for inferences in absence of missing data is also valid when missing data are MCAR, e.g.

- Analysis based on all available data
- Analysis based on the so-called "completers"

Missing at random (MAR)

MAR: probability that responses are missing depends on the set of observed responses, but is unrelated to the specific missing values that, in principle, should have been obtained.

MAR: probability that responses are missing depends on $Y^{(o)}$, but is conditionally independent of $Y^{(m)}$.

Note 1: If subjects are stratified on the basis of similar values for the responses that have been observed, then within strata missingness is simply the result of a chance mechanism unrelated to unobserved responses.

Note 2: The "completers" are a biased sample from the target population

Features of MAR

The means, variances and covariances are NOT preserved

If

$E(Y_i) = X_i \beta$ with complete data $Y_i - n \times 1$, $X_i - n \times p$

then in general

$E(Y_i) \neq X_i \beta$ $Y_i - n_i \times 1$, $X_i - n_i \times p$

and

$Cov(Y_i) \neq \Sigma_i$

This implies that sample means, variances, and covariances based on either the "completers" or the available data are biased estimates of the corresponding parameters in the target population.

Features of MAR

However, the likelihood is preserved.

For example, in the linear models for longitudinal data the appropriate likelihood assumes

$$Y_i \sim N(X_i\beta,\Sigma_i).$$

Because missingness only depends on observed data, likelihood factors into one piece depending on (β,Σ_i), another depending on Y_i and missingness indicators.

Valid inferences for (β,Σ_i) are obtained by maximizing the first piece (and "ignoring" the second piece) of the likelihood.

Note: Observed responses are not necessarily normally distributed.

Estimation under MAR

ML estimation (e.g., function lme() in R) of β is valid when data are MAR provided the multivariate normal distribution has been correctly specified.

This requires correct specification of not only the model for the mean response, but also the model for the covariance among the responses.

In a sense, ML estimation allows the missing values to be validly "predicted" or "imputed" using the observed data and a correct model for the joint distribution of the responses.

Not Missing at Random (NMAR)

NMAR: probability that responses are missing is related to the specific values that should have been obtained.

An NMAR mechanism is often referred to as “non-ignorable” missingness.

Challenging problem that requires modeling of missing data mechanism; moreover, specific model chosen can drive results of analysis.

Sensitivity analyses is recommended.

Dropout

Longitudinal studies often suffer from problem of attrition; i.e., some individuals "drop out" of the study prematurely.

This is when an individual is observed from baseline up until a certain point in time, thereafter no more measurements are made.

Term dropout refers to special case where

if Y_{ik} is missing, then Y_{ik+1}, ..., Y_{in} are also missing.

This gives rise to so-called **"monotone"** missing data pattern.

Possible reasons for dropout in clinical trials

1. Recovery

2. Lack of improvement or failure

3. Undesirable side effects

4. External reasons unrelated to specific treatment or outcome

5. Death

Examples

In clinical trials, monotone missing data can arise for a variety of reasons:

a) Late entrants:

If the study has staggered entry, at any interim analysis some individuals may have only partial response data. Usually, this sort of missing data does not introduce any bias.

b) Dropout:

Individuals may drop out of a clinical trial because of side effects or lack of efficacy. Usually, this type of missing data is of concern, especially if dropout is due to lack of efficacy.

Dropout due to lack of efficacy suggests that those who drop out may be primarily those who are not doing well.

Dropout due to side effects may or may not be a problem, depending upon the relationship between side effects and the outcome of interest.

Methods of Handling Missing Data

- Complete Case Analysis
- Available data analysis
- Imputation
- Multiple imputation
- Last value carried forward (LVCF)
- Model-Based imputation
- Weighting methods

Complete-case analysis

These methods omit all cases with missing values at any measurement occasion.

Exclude all data from the analysis on any subject who drops out.

That is, a so-called “complete-case” analysis can be performed by excluding any subjects that do not have data at all intended measurement occasions.

This method is very problematic and is rarely an acceptable approach to the analysis.

It will yield unbiased estimates of mean response trends only when dropout is MCAR.

Even when MCAR assumption is tenable, complete-case analysis can be very inefficient.

Available data analysis

General term that refers to a wide collection of techniques that can readily incorporate vectors of repeated measures of unequal length in the analysis.

Standard applications of GLS are available-data methods.

In general, available-data methods are more efficient than complete-case methods.

Drawbacks of available-data methods:

(i) Sample base of cases changes over measurement occasions.

(ii) Pairwise available-data estimates of correlations can lie outside (-1, 1).

(iii) Many available-data methods yield biased estimates of mean response trends unless dropout is MCAR.

Imputation

In this approach, we substitute or fill-in the values that were not recorded with imputed values.

Once a filled-in data set has been constructed, standard methods for complete data can be applied.

Validity of method depends on how imputation is done.

Methods that rely on just a single imputation fail to acknowledge the uncertainty inherent in the imputation of the unobserved responses.

“Multiple imputation” circumvents this difficulty.

Multiple imputation

Missing values are replaced by a set of “m” plausible values, thereby acknowledging uncertainty about what values to impute.

Typically, a small number of imputations, for instance, $5 \leq m \leq 10$, is sufficient.

The “m” filled-in data sets produce m different sets of parameter estimates and their standard errors.

These are then combined to provide a single estimate of the parameters of interest, together with standard errors that reflect the uncertainty inherent in the imputation.

Last value carried forward (LVCF)

One widely used imputation method, especially in clinical trials, is LVCF.

In the past, regulatory agencies such as FDA seem to encourage the continuing use of LVCF.

However, LVCF makes a strong, and often very unrealistic, assumption that the responses following dropout remain constant at the last observed value prior to dropout.

There appears to be some statistical folklore that LVCF yields a conservative estimate of the comparison of an active treatment versus the control.

This is a gross misconception!

Except for very rare cases, the use of LVCF as a method for handling dropout is **NOT** recommended.

Last value carried forward (LVCF)

Variations on the LVCF theme include:

- baseline value carried forward and
- worst value carried forward.

Imputation methods based on drawing values of missing responses from the conditional distribution of the missing responses given the observed responses have a much firmer theoretical foundation.

Then subsequent analyses of the observed and imputed data are valid when dropouts are MAR (or MCAR). Furthermore, multiple imputation ensures that the uncertainty is properly accounted for.

Model-based imputation

There is a related form of "imputation" where missing responses are implicitly imputed by modeling joint distribution of Y_i, $f(Y_i | X_i)$.

When dropout is MCAR or MAR, likelihood-based methods can be used based solely on the marginal distribution of the observed data.

In a certain sense, the missing values are validly predicted by the observed data via the model for the conditional mean of the missing responses given the observed responses (and covariates).

However, likelihood-based approaches require model for $f(Y_i | X_i)$ to be correctly specified (e.g., any misspecification of the covariance will, in general, yield biased estimates of the mean response trend).

Weighting methods

In weighting methods, under-representation of certain response profiles in the observed data is taken into account and corrected.

These approaches are often called "**propensity weighted**" or "**inverse probability weighted**" methods.

Basic Idea: Base estimation on the observed responses but weigh them to account for the probability of remaining in the study.

Intuition: Each subject's contribution to the weighted analysis is replicated to count for herself and for those subjects with the same history of responses and covariates, but who dropped out.

Weighting methods

Propensities for dropout can be estimated as a function of observed responses prior to dropout and covariates.

Inverse probability weighted methods were first proposed in sample survey literature, where the weights are known.

In contrast, with dropout the weights are not known, but must be estimated from the observed data.

In general, weighting methods are valid provided model that produces the estimated weights is correctly specified.

Summary

In longitudinal studies **missing data** are the **rule** NOT the exception.

Missing data have two important implications:

(i) loss of information, and

(ii) validity of analysis.

The loss of information is directly related to the amount of missing data; it will lead to reduced precision (e.g., larger SEs, wider CIs) and reduced statistical power (e.g., larger p-values).

The validity of the analysis depends on assumptions about the missing data mechanism.

Summary

Likelihood-based methods (e.g., function lme() in R) are valid under MAR or MCAR.

The distinction between MAR and MCAR determines the appropriateness of ML estimation under the assumption of normality versus GLS estimation without requiring distributional assumptions.

With complete data or data MCAR, normality assumption is not required.

With data MAR, normality assumption is required and correct models for both the mean response and the covariance.