Missing Data

MSc Further Statistical Methods, Lecture 5
Hilary Term 2005

Steffen Lauritzen, University of Oxford; February 7, 2005

Missing data problems

case	A	B	C	D	E	F
1	a_1	b_1	$*$	d_1	e_1	*
2	a_2	$*$	c_2	d_2	e_2	$*$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	a_n	b_n	c_n	$*$	$*$	$*$

$*$ or *NA* denotes values that are *missing*, i.e. non-observed.

Examples of missingness

- non-reply in surveys, "missing" don't know, essentially an additional state for the variable in question
- recording error
- variable out of range
- just not recorded (e.g. too expensive)

Different types of missingness demand different treatment.

Notation for missingness

Data matrix Y, *missing data matrix* $M = \{M_{ij}\}$:

$$M_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is missing} \\ 0 & \text{if } Y_{ij} \text{ is observed.} \end{cases}$$

Convenient to introduce the notation $Y = (Y_{\mathrm{obs}}, Y_{\mathrm{mis}})$, where Y_{mis} are conceptual and denote the data that were not observed.

Patterns of missingness

Univariate: $M_{ij} = 0$ unless $j = j^*$, e.g. an unmeasured response

Multivariate: $M_{ij} = 0$ unless $j \in J \subset V$, as above, just with multivariate response, e.g. in surveys

Monotone: There is an ordering of V so $M_{ik} = 0$ implies $M_{ij} = 0$ for $j < k$, e.g. drop-out in longitudinal studies.

Disjoint: Two subsets of variables never observed together. Controversial. Appears in Rubin's causal model.

General: none of the above. Haphazardly scattered missing values.

Latent: A certain variable is never observed. Maybe it is even unobservable.

Methods for analysis tend to get increasingly complex as we go down the list.

Methods for dealing with missing data

Complete case analysis: analyse only cases where all variables are observed. Can be adequate if most cases are present, but will generally give serious biases in the analysis. In survey's, for example, this corresponds to making inference about the population of responders, not the full population;

Weighting methods. For example, if a population total $\mu = \mathbf{E}(Y)$ should be estimated and unit i has been selected with probability π_i a standard method is the *Horwitz–Thompson estimator*

$$\hat{\mu} = \frac{\sum \frac{Y_i}{\pi_i}}{\sum \frac{1}{\pi_i}}.$$

To correct for non-response, one could let ρ_i be the response-probability, estimate this in some way as $\hat{\rho}_i$ and then let

$$\tilde{\mu} = \frac{\sum \frac{Y_i}{\pi_i \hat{\rho}_i}}{\sum \frac{1}{\pi_i \hat{\rho}_i}}.$$

Imputation methods: Find ways of estimating the values of the unobserved values as $\hat{Y}_{\mathrm{mis}}$, then proceed as if there were complete data. Without care, this can give misleading results, in particular because the "sample size" can be grossly overestimated.

Model-based likelihood methods: Model the missing data mechanism and then proceed to make a proper likelihood-based analysis, either via the method of maximum-likelihood or using Bayesian methods. This appears to be the most sensible way.

Typically this approach was not computationally feasible in the past, but modern algorithms and computers have changed things completely. Ironically, the efficient algorithms are indeed based upon imputation of missing values, but with proper corrections resulting.

Mechanisms of missingness

The data are *missing completely at random*, MCAR, if

$$f(M \mid Y, \theta) = f(M \mid \theta), \text{ i.e. } M \perp\!\!\!\perp Y \mid \theta.$$

Heuristically, the values of Y have themselves no influence on the missingness. Example is recording error, latent variables, and variables that are missing *by design* (e.g. measuring certain values only for the first m out of n cases). Beware: it may be counterintuitive that *missing by design is MCAR*.

The data are *missing at random*, MAR, if

$$f(M \mid Y, \theta) = f(M \mid Y_{\mathrm{obs}}, \theta), \text{ i.e. } M \perp\!\!\!\perp Y_{\mathrm{mis}} \mid (Y_{\mathrm{obs}}, \theta).$$

Heuristically, only the observed values of Y have influence on the missingness. By design, e.g. if individuals with certain characteristics of Y_{obs} are not included in part of study (where Y_{mis} is measured).

The data are *not missing at random*, NMAR, in all other cases.

For example, if certain values of Y cannot be recorded when they are out of range, e.g. in survival analysis.

The classifications above of the mechanism of missingness lead again to increasingly complex analyses.

It is not clear than the notion MCAR is helpful, but MAR is. Note that *if data are MCAR, they are also MAR.*

Likelihood-based methods

The most convincing treatment of missing data problems seems to be via modelling the missing data mechanism, i.e. *by considering the missing data matrix M as an explicit part of the data.*

The likelihood function then takes the form

$$L(\theta \,|\, M, y_{\text{obs}}) \propto \int f(M, y_{\text{obs}}, y_{\text{mis}} \,|\, \theta)\, dy_{\text{mis}} \qquad (1)$$

with

$$f(M, y_{\text{obs}}, y_{\text{mis}} \,|\, \theta) \propto L_{\text{mis}}(\theta) f(y_{\text{obs}}, y_{\text{mis}} \,|\, \theta), \qquad (2)$$

where the term $L_{\text{mis}}(\theta) \propto f(M \,|\, y_{\text{obs}}, y_{\text{mis}}, \theta)$ is based on an explicit model for the missing data mechanism.

Ignoring the missing data mechanism

The likelihood function *ignoring the missing data mechanism* is

$$L_{\text{ign}}(\theta \,|\, y_{\text{obs}}) \propto f(y_{\text{obs}} \,|\, \theta) = \int f(y_{\text{obs}}, y_{\text{mis}} \,|\, \theta)\, dy_{\text{mis}}. \quad (3)$$

When is $L \propto L_{\text{ign}}$ so the missing data mechanism can be ignored for further analysis? We will show this is true under the *following conditions:*

1. The data are *MAR*;
2. The parameters η governing the missingness are *separate* from parameters of interest ψ, so that information about the value of one does not restrict the other.

Ignorable missingness

If data are MAR and the missingness parameter is separate from the parameter of interest, we have $\theta = (\eta, \psi)$ and

$$L_{\mathrm{mis}}(\theta) = L_{\mathrm{mis}}(\eta) \propto f(M \mid y_{\mathrm{obs}}, y_{\mathrm{mis}}, \eta) = f(M \mid y_{\mathrm{obs}}, \eta)$$

Hence, the factor L_{mis} is constant in (2) and can be taken outside in the integral in (1) so that, combining with (3) and

$$f(y_{\mathrm{obs}}, y_{\mathrm{mis}} \mid \theta) = f(y_{\mathrm{obs}}, y_{\mathrm{mis}} \mid \psi)$$

we get

$$L(\theta \mid M, y_{\mathrm{obs}}) \propto L_{\mathrm{mis}}(\eta) L_{\mathrm{ign}}(\psi \mid y_{\mathrm{obs}})$$

which shows that the missingness mechanism can be ignored when concerned with likelihood inference about ψ.

For a Bayesian analysis the parameters must in addition be *independent w.r.t. the prior:*

$$f(\eta, \psi) = f(\eta)f(\psi).$$

If the data are *NMAR* or the parameters are not separate, then the missing data mechanism cannot be ignored, care must be taken to model the mechanism $f(M \mid y_{\text{obs}}, y_{\text{mis}}, \theta)$ and the corresponding likelihood term must be properly included in the analysis.