Statistical Methods

Missing Data

http://www.stats.ox.ac.uk/~snijders/sm.htm

Tom A.B. Snijders

University of Oxford

November, 2011

Literature:

Joseph L. Schafer and John W. Graham,
Missing Data: Our View of the State of the Art.
Psychological Methods, 7 (2002), 147–177.

Useful further literature:

John W. Graham,
Missing Data Analysis: Making It Work in the Real World.
Annual Reviews of Psychology, 60 (2009), 349–376.

Andrew Gelman & Jennifer Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. CUP, 2007.

Chapter 25: Missing-data imputation.

Missing data ...?

Statistics mainly is about available data!

Missing data has been a subject of ad hoc methods,
and many practitioners approach missing data issue
in a naive way.

Better and principled methods exist,
especially starting with the work by Donald Rubin in the 1970s.

Missing data ...?

Statistics mainly is about available data!

Missing data has been a subject of ad hoc methods,
and many practitioners approach missing data issue
in a naive way.

Better and principled methods exist,
especially starting with the work by Donald Rubin in the 1970s.

First some basic definitions.

Missingness patterns

(p. 151) (*page numbers refer to Schafer & Graham, 2002*)
Suppose the data set is a $n \times p$ matrix $Y = (Y_{ij})$
and the *missing data indicator matrix* $M = (M_{ij})$ indicates whether data point Y_{ij} is present ($M_{ij} = 0$) or missing ($M_{ij} = 1$).
The missingness mechanism is supposed to be modeled by a set of probability distributions $\mathcal{L}(M \mid \psi)$, with parameter ψ.

MCAR If missingness is independent of data:
$\mathrm{P}\{M = m \mid Y, \psi\} = \mathrm{P}\{M = m \mid \psi\}$
for all m, ψ,
then the missingness model is said to be
missing completely at random.

Example: randomly failing recording apparatus.

Represent by $Y = (Y_{\text{obs}}, Y_{\text{mis}})$
the partition of the data into observed and missing data
where Y_{ij} is part of Y_{mis} if and only if $M_{ij} = 1$,
and part of Y_{obs} otherwise.

MAR If missingness depends only on observed data,
and not on missing data :
$\mathrm{P}\{M = m \mid Y, \psi\} = \mathrm{P}\{M = m \mid Y_{\text{obs}}, \psi\}$
for all m, ψ,
then the missingness model is said to be
missing at random.

Example: repeated measures of ill patients where observation stops at recovery, and recovery is part of the data.

If missingness depends on unobserved data
(i.e., MAR does not hold), then the
missingness model is missing not at random, MNAR.

Examples: clinical trials,
where patients drop out because they are too ill, or recovered,
or because they do not wish to continue treatment,
without this being recorded;
another example is censoring on unobserved variables.

Naive ways of dealing with missingness
usually assume MCAR without thinking.

Large improvements of handling missing data can be achieved
by using appropriate methods
and exploiting the MAR assumption.
(The phrase “missing at random” can be misleading,
because we might think it means MCAR.)

For MNAR, we are in difficulties; but sometimes it is possible
to make meaningful models of the missingness mechanism;
this will be based on partially untestable assumptions,
and a sensitivity analysis often is appropriate.

This lecture is about MAR methods.

Example

Suppose that $p = 2$, Y_{i1} is age and Y_{i2} is income,
and Y_{i1} is completely observed.
Define $M_i = 1$ if income is observed for respondent i.
Define $Y_i = (Y_{i1}, Y_{i2})$ and suppose that (Y_i, M_i) are i.i.d.

Example

Suppose that $p = 2$, Y_{i1} is age and Y_{i2} is income,
and Y_{i1} is completely observed.
Define $M_i = 1$ if income is observed for respondent i.
Define $Y_i = (Y_{i1}, Y_{i2})$ and suppose that (Y_i, M_i) are i.i.d.

If M_i and Y_i are independent, i.e., every respondent
has the same probability of income being unobserved,
then MCAR holds.

Example

Suppose that $p = 2$, Y_{i1} is age and Y_{i2} is income,
and Y_{i1} is completely observed.
Define $M_i = 1$ if income is observed for respondent i.
Define $Y_i = (Y_{i1}, Y_{i2})$ and suppose that (Y_i, M_i) are i.i.d.

If M_i and Y_i are independent, i.e., every respondent
has the same probability of income being unobserved,
then MCAR holds.

If M_i depends on age but not on unobserved income,
then MAR holds – given that we observed age!

Example

Suppose that $p = 2$, Y_{i1} is age and Y_{i2} is income,
and Y_{i1} is completely observed.
Define $M_i = 1$ if income is observed for respondent i.
Define $Y_i = (Y_{i1}, Y_{i2})$ and suppose that (Y_i, M_i) are i.i.d.

If M_i and Y_i are independent, i.e., every respondent
has the same probability of income being unobserved,
then MCAR holds.

If M_i depends on age but not on unobserved income,
then MAR holds – given that we observed age!

If M_i depends on the unobserved income, then MNAR holds
– and we are in trouble.

Practical issues

Practical issues

$\Rightarrow$ In the design of the data collection,
take care that missingness is avoided/minimized;
but *missingness by design* is allowed.

Practical issues

⇒ In the design of the data collection,
take care that missingness is avoided/minimized;
but *missingness by design* is allowed.

⇒ If missingness is unavoidable, then
collect variables that are predictive of missingness,
and of the unobserved values.
This helps for the plausibility of the MAR assumption.

Practical issues

- ⇒ In the design of the data collection,
  take care that missingness is avoided/minimized;
  but *missingness by design* is allowed.
- ⇒ If missingness is unavoidable, then
  collect variables that are predictive of missingness,
  and of the unobserved values.
  This helps for the plausibility of the MAR assumption.
- ⇒ If there are missing values,
  try to understand how they arose,
  and describe their frequency and patterns.

Practical issues

- ⇒ In the design of the data collection,
  take care that missingness is avoided/minimized;
  but *missingness by design* is allowed.
- ⇒ If missingness is unavoidable, then
  collect variables that are predictive of missingness,
  and of the unobserved values.
  This helps for the plausibility of the MAR assumption.
- ⇒ If there are missing values,
  try to understand how they arose,
  and describe their frequency and patterns.
- ⇒ Use adequate procedures dealing with the missingness.

When there are missing data, part of the descriptive analysis of the data set always should be an informal investigation of the missingness patterns:

how many missing values are there;
how are they clustered for certain variables;
are there systematic differences
– with respect to observed variables –
between cases with and without missing values;
it may be meaningful to make a logistic regression analysis of missingness of important variables,
with observed variables as predictors.

Knowing what is predictive for missingness within the available data can help understanding the processes leading to missingness.

Regression analysis under MAR

Suppose that we are interested in a regression analysis with predictor variables X and dependent variable Y.

If X is completely observed and missings in Y are MAR, and also the parameters of the missingness model are unrelated to the parameters of the conditional distribution $\mathcal{L}(Y \mid X)$, then regression analysis on the complete cases is a valid approach.

This has the advantage that no assumptions are necessary about the joint distribution of the variables in X.

Naive approach: Complete cases analysis

(p. 155)
An easy way out is to use only complete cases ($\sum_j M_{ij} = 0$)
and carry out the analysis as if there were no missing data.
This is also called *listwise deletion*.
In many programs (including R package lm) it is the default.

This is unbiased but inefficient under MCAR,
but is likely to lead to bias under MAR.

The example in Schafer & Graham (p. 152-153)
of estimating $\mathrm{E}\, Y_{i2}$ under MAR
shows the precise amount of bias in Table 2.

Another example in script `misdat.r`.

Adequate methods

There are various good methods for analyzing incomplete data under the MAR assumption.

The most frequently used ones are

$\Rightarrow$ Full Information Maximum Likelihood:
based on the likelihood function for the incomplete data;
if MAR is valid, the missingness mechanism
does not have to be explicitly modeled.
Example in script `misdat.r`.

Adequate methods

There are various good methods for analyzing incomplete data under the MAR assumption.

The most frequently used ones are

- ⇒ Full Information Maximum Likelihood:
  based on the likelihood function for the incomplete data;
  if MAR is valid, the missingness mechanism
  does not have to be explicitly modeled.
  Example in script `misdat.r`.
- ⇒ Multiple Imputation.
  This is more often possible than FIML,
  and will be explained here.

Imputation

(p. 158)
Since most statistical methods and algorithms
are defined for complete cases,
a tempting approach is to impute, i.e., fill in, missing values
and then use a method for complete cases.

Many methods for imputation have been proposed;
some are better than others.

Unconditional mean imputation is easy,
but usually not a good idea:
underestimation of variability;
distortion of relations with other variables.

Imputing random draws

(p. 159)
The answer to the problem of underestimated variability is to add random variability.

For example, if Y has a multivariate normal distribution then for each case i we may substitute missings by *random draws from the conditional normal distribution of the missing data, given the observed data*.

Imputing random draws

(p. 159)
The answer to the problem of underestimated variability
is to add random variability.

For example, if Y has a multivariate normal distribution
then for each case i we may substitute missings
by *random draws from the conditional normal distribution of the missing data, given the observed data*.

If MAR holds, then this will be a reasonable procedure;
the multivariate normality assumption is not very critical
if the number of missing values is not too high.

But the answer is random because of the added randomness.
Doing it multiple times is better: multiple imputation.

Multiple imputation

(p. 165)

For a given incomplete data set,
the missing data is imputed independently D times
by drawing from the conditional distribution
of the missing data given the observed data.

This leads to D complete data sets,
that differ only with respect to the imputed values.

Multiple imputation

(p. 165)

For a given incomplete data set,
the missing data is imputed independently D times
by drawing from the conditional distribution
of the missing data given the observed data.

This leads to D complete data sets,
that differ only with respect to the imputed values.

For each complete data set the desired analysis is executed;
standard errors of parameters are a combination
of the within-data set standard errors,
and the variability of estimates between the data sets.

This between-imputed data sets variability will be larger,
as the amount of missing data is larger.

There are two questions in practice:

1. How to impute?
   (Draw from the conditional distribution
   of missings given observed)
2. How to combine the D data sets?

How to impute

Imputation requires that the joint distribution of the data is known, but this still is to be estimated...

Here the guidelines are

1. Separate the *imputation model* from the *analysis model*; the imputation model can include more observed variables than the analysis model,
2. the imputation model must represent the associations between the variables well. Often, sampling under the assumption of multivariate normality is reasonable. Example: scripts `misdat.r` and `missingdat.r`.

Another good imputation method is *data augmentation*, a Bayesian technique implemented usually by MCMC. Strong point: propagation of parameter uncertainty.

How to combine

The parameter of interest is denoted γ.

Suppose that the d'th randomly imputed data set leads to estimates $\hat{\gamma}_d$ and estimated variances W_d ('Within'),

$$W_d = \text{var}\{\hat{\gamma}_d \mid \text{ data set } d\} .$$

Note that W_d underestimates true uncertainty,
because it treats imputed data as real data.

The formulae below indicate how the D results are put together,
for the case that the estimated parameter γ is a scalar.

They were developed by Donald Rubin (1987).

The combined estimate is the average

$$\bar{\gamma}_D = \frac{1}{D}\sum_{d=1}^{D}\hat{\gamma}_d \, .$$

Compute the average within-imputation variance

$$\overline{W}_D = \frac{1}{D}\sum_{d=1}^{D} W_d \, ,$$

and the between-imputation variance

$$B_D = \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\gamma}_d - \bar{\gamma}_D\right)^2 \, .$$

Estimated total variability for $\bar{\gamma}_D$ is

$$T_D = \widehat{\text{var}}\left(\bar{\gamma}_D\right) = \overline{W}_D + \frac{D+1}{D}B_D \, .$$

This shows that

$$\hat{\pi}_D = \frac{T_D - \overline{W}_D}{T_D} = \frac{(1 + D^{-1})B_D}{\overline{W}_D + (1 + D^{-1})B_D}$$

is the estimated extra variability due to missingness of data; this is called the *fraction of missingness*.

Hypothesis tests can be based on the reference distribution

$$\frac{\bar{\gamma}_D - \gamma}{\sqrt{T_D}} \sim t_\nu$$

where the number of degrees of freedom is

$$\nu_D = (D-1)\Big(1 + \frac{\overline{W}}{(1 + D^{-1})B_D}\Big)^2 .$$

This procedure can be followed already for $D = 20$; larger fractions of missingness require larger D.

Example

See script `missingdat.r` for an elaborate example.

Some R packages useful for inference with missing data.

Amelia II	Bootstrap EM imputation.
cat	Missing data methods for categorical data.
mi	Missing Data Imputation and Model Checking.
mice	Multiple Imputation and generalized linear regression by Chained Equations.
mix	Missing data methods for mixed categorical & continuous data.
mlmmm	Estimation for mixed linear models with missing data.
mvnmle	MLE for multivariate normal with missing data.
norm	Estimation and imputation for multivariate normal data with missings.
VIM	Visualization and Imputation of Missing Values.

There are many more!