MISSING VALUES

ARTHUR ALLIGNOL

INTRODUCTION

INTRODUCTION

Missing data are ubiquitous

- Missing data for some variables (e.g., a test not done...)
- for some cases (e.g., record lost)

Bad handling of missing data can lead to

- Blurred effect estimates
- and/or biased effect estimates

TYPES OF MISSING DATA

TYPES OF MISSING DATA

Suppose that the data set is a $n \times p$ matrix $Y = (Y_{ij})$

Let $M = (M_{ij})$ be a matrix indicating whether the data point Y_{ij} is

- present: $M_{ij} = 0$
- missing: $M_{ij} = 1$

The missingness mechanism is supposed to be modelled by a set of probability distributions $\mathcal{L}(M|\psi)$

MISSING COMPLETELY AT RANDOM

if missingness is independent of the data,

$$\Pr(M = m|Y, \psi) = \Pr(M = m|\psi) \quad \text{for all } m \text{ and } \psi,$$

then the missingness model is said to be *Missing Completely At Random* (MCAR)

Because the mechanism of missingness is only due to chance, deleting MCAR data will not bias the results, but the effective sample size is lowered (loss of efficiency)

MISSING AT RANDOM

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

If the missingness only depends on the observed data,

$$\Pr(M = m|Y, \psi) = \Pr(M = m|Y_{\text{obs}}, \psi) \quad \text{for all } m \text{ and } \psi,$$

then the missingness is said to be *Missing At Random* (MAR)

MISSING NOT AT RANDOM

If the missingness depends on unobserved data (i.e., MAR does not hold) then the missingness model is said to be *Missing Not At Random* (MNAR)

That's the most annoying situation...

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.

INTRODUCTION

Naive ways of dealing with missingness usually assume MCAR without thinking

Large improvements in handling missing data can be had by using appropriate methods and exploiting the MAR assumption

MNAR is another beast...

We focus on MAR data in this lecture

HOW NOT TO HANDLE MISSING DATA

BAD WAYS TO HANDLE MISSING DATA

Deletion methods

- *Listwise deletion*: delete the entire observation if any values are missing

  That's what, e.g., `lm()`, `glm()` does if either outcome variable or dependent variables are missing

  For longitudinal models (e.g., `lmer`), listwise deletion only if explanatory variables are missing

- *Pairwise deletion*: delete a pair of observations if either of the values are missing

If the missing data are *MCAR*, deletions will result in *unbiased* estimates

But:

- Reduction in statistical power if MCAR
- Biased estimates if MAR or MNAR

BAD WAYS TO HANDLE MISSING DATA

Single imputation methods

Single imputation methods replace missing data with some type of data

- *Single:* One value used
- *Imputation:* Replace missing data with value

One can then use the entire data set for fitting the model

BUT biased parameter estimates and standard errors even if the missing data are MCAR

BAD WAYS TO HANDLE MISSING DATA

Last Observation Carried Forward (LOCF; for longitudinal data)

The idea of LOCF is to replace observations that dropped out by the last available value

- That assumes that the variable do not change after drop out

  Thought to be conservative (assuming a monotone beneficial effect of a treatment)

  The other way around, it can exaggerate group differences

BAD WAYS TO HANDLE MISSING DATA

Summary

All the methods presented so far are bad for handling missing values

- They all lead to too small standard errors (→ inflation of the type I error)
- Parameter estimates likely to be biased (in either direction)
  - Exception is deletion under MCAR

PREFERRED METHODS FOR DEALING WITH MISSING VALUES

MULTIPLE IMPUTATION

Let Q be the population quantity of interest. For instance a regression coefficient

If all the data have been observed, estimates and inferences for Q would have been based on the complete-data posterior density

$$f(Q|Y_{\text{obs}}, Y_{\text{mis}})$$

As is Y_{mis} is not observed, inferences have to be based on the actual posterior density

$$f(Q|Y_{\text{obs}}) = \int f(Q|Y_{\text{obs}}, Y_{\text{mis}}) f(Y_{\text{mis}}|Y_{\text{obs}}) \mathrm{d}Y_{\text{mis}}$$

MULTIPLE IMPUTATION

$$f(Q|Y_{\text{obs}}) = \int f(Q|Y_{\text{obs}}, Y_{\text{mis}}) f(Y_{\text{mis}}|Y_{\text{obs}}) \mathrm{d}Y_{\text{mis}}$$

The actual posterior density of Q can be obtained by averaging the complete posterior density over the posterior predictive distribution of Y_{mis}

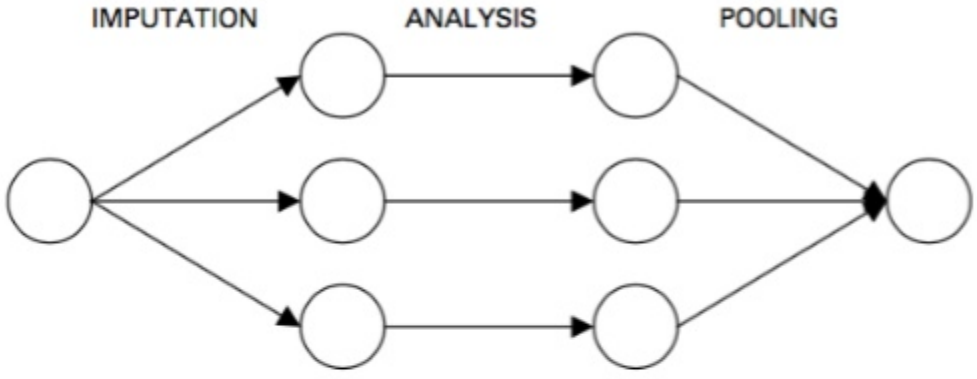
In a nutshell, the idea behind MI is to independently draw multiple times from $f(Y_{\text{mis}}|Y_{\text{obs}})$.

Then multiple imputation allows approximating the above equation by separately analysing each data set completed by imputation and then combining the results

MULTIPLE IMPUTATION

The idea of multiple imputation is

1. Impute missing values by drawing from the observed values
2. Repeat the process several times
3. Average the results in order to get an estimate with a measure of uncertainty that accounts for the uncertainty due to imputation



$f(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}})$ Impute based on missing data model

$f(Q|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}})$ Outcome model using complete data

$\int \mathrm{d}\mathbf{Y}_{\text{mis}}$ "Integrate" over imputed datasets

$f(Q|\mathbf{Y}_{\text{obs}})$ What you get

$$f(Q|\mathbf{Y}_{\text{obs}}) = \int f(Q|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}) f(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}) \mathrm{d}\mathbf{Y}_{\text{mis}}$$

Little 2002

MULTIPLE IMPUTATION — A SIMPLE EXAMPLE

```
set.seed(10)
x <- c(sample(1:10, 7, TRUE), rep(NA, 3))
x
```

```
 [1]  6  4  5  7  1  3  3 NA NA NA
```

MULTIPLE IMPUTATION — A SIMPLE EXAMPLE

- Compute mean using case deletion

```
mean(x, na.rm = TRUE)
```

```
[1] 4.142857
```

- Standard error

```
sd(x, na.rm = TRUE)/sqrt(sum(!is.na(x)))
```

```
[1] 0.7693093
```

- Single mean imputation

```
x_imp <- c(x[!is.na(x)], rep(mean(x, na.rm = TRUE), 3))
sd(x_imp)/sqrt(10)
```

```
[1] 0.5255383
```

MULTIPLE IMPUTATION — A SIMPLE EXAMPLE

- Now let's impute several times to generate a list of imputed vectors:

```
imp <- replicate(15, c(x[!is.na(x)], sample(x[!is.na(x)], 3, TRUE)), simplify = FALSE)
imp
```

```
[[1]]
 [1] 6 4 5 7 1 3 3 4 1 7

[[2]]
 [1] 6 4 5 7 1 3 3 1 7 6

[[3]]
 [1] 6 4 5 7 1 3 3 1 5 7

[[4]]
 [1] 6 4 5 7 1 3 3 6 4 5

[[5]]
 [1] 6 4 5 7 1 3 3 3 3 1

[[6]]
 [1] 6 4 5 7 1 3 3 3 5 5

[[7]]
 [1] 6 4 5 7 1 3 3 1 3 4

[[8]]
 [1] 6 4 5 7 1 3 3 3 5 7
```

MULTIPLE IMPUTATION — A SIMPLE EXAMPLE

- Mean for each imputed vector

```
means <- sapply(imp, mean)
means
```

```
 [1] 4.1 4.3 4.2 4.4 3.6 4.2 3.7 4.4 4.2 4.0 4.0 4.3 4.0 4.8 4.0
```

- Average over the imputed vectors

```
grandm <- mean(means)
grandm
```

```
[1] 4.146667
```

MULTIPLE IMPUTATION — A SIMPLE EXAMPLE

- Obtain standard errors

  We need to combine the within- and between-imputation variance

```
ses <- sapply(imp, sd)/sqrt(10)
within <- mean(ses)
between <- sum((means - grandm)^2)/(length(imp) - 1)

grandvar <- within + ((1 + (1/length(imp))) * between)
grandse <- sqrt(grandvar)
grandse
```

```
[1] 0.8387083
```

Note that the SE is bigger than that of the complete case analysis. That's actually a good thing because we need to take into account the uncertainty due to the imputation

IMPUTING THE MISSING DATA

Let the hypothetically complete data be a partially observed random sample from the multivariate distribution $P(Y|\theta)$, with θ a vector of unknown parameters.

One approach is to assume that $P(Y|\theta)$ can be modelled by a joint multivariate distribution, e.g., multivariate normal. Then the imputed values can be drawn from the corresponding predictive distribution.

Although theoretically sound this approach is limited if Y contains a mix of continuous and binary variables

FULLY CONDITIONAL SPECIFICATION

The idea of the fully conditional specification is to obtain a posterior distribution of θ by sampling iteratively from conditional distribution of the form

$$p(Y_1|Y_{-1},\theta_1)$$
$$\vdots$$
$$p(Y_p|Y_{-p},\theta_p)$$

$\theta_1,\cdots,\theta_p$ are specific to the respective conditional densities

MICE

Multiple Imputation by Chained Equations is a popular implementation of the fully conditional specification. The algorithm works as follows

For the first iteration $Y_j^{(0)}$, random values of the observed Y_j are drawn

Now assume that the algorithm is at the t-th iteration. The imputations at the next iteration are random draws from

$$
\begin{aligned}
\theta_1^{*(t)} &\sim p(\theta_1 | Y_{1;\text{obs}}, Y_{-1}^{(t-1)}) \\
Y_1^{(t)} &\sim p(Y_1 | Y_{1;\text{obs}}, Y_{-1}^{(t-1)}; \theta_1^{*(t)}) \\
&\vdots \\
\theta_p{*}(t) &\sim p(\theta_p | Y_{p;\text{obs}}, Y_{-p}^{(t)}) \\
Y_p^{(t)} &\sim p(Y_p | Y_{p;\text{obs}}, Y_{-p}^{(t)}; \theta_p^{*(t)})
\end{aligned}
$$

THE IMPUTATION MODELS

We need to decide on

- The structural form of the imputation model; usually steered by the scale of the variable to impute
- Which variables to include: As many relevant variables as possible. Note that the *outcome variable* as well as the *dependent variables* included in the *analysis model* should be included
- Order in which variables should be imputed
- Number of iterations of the algorithm. It's usually fast, but convergence should be monitored
- Number of imputed data sets m. Too low leads to undercoverage. Some argues that $m = 5$ is enough; others say $m \geq 50$

FULLY CONDITIONAL SPECIFICATION

Pros:

FCS is extremely flexible

- One model per variable to impute permits to avoid implausible values (e.g., gender equal to 0.7)
- There is ways to specify how some variables are related in order to avoid pregnant fathers
- Passive imputation: Impute height and weight. The software takes care of BMI
- Shown to work quite well in practice
- Can be extended to MNAR by specifying the right missing data mechanism

Cons

- Theoretical rational not well established
- Appropriateness of the algorithm mostly demonstrated through simulation studies. Some work might still be needed in order to identify the boundaries at which the algorithm breaks down

ANALYSING MULTIPLY IMPUTED DATA

Let $\hat{Q}_i$ be the estimate from the i-th data set ad S_i the corresponding variance.

The combined estimate of Q is

$$\bar{Q} = \frac{1}{m} \sum_{i=1}^{m} \hat{Q}_i$$

The combined variance depends on the within-imputation variance

$$\bar{S} = \frac{1}{m} \sum_{i=1}^{m} S_i$$

and the between-imputation variance

$$B = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{Q}_i - \bar{Q})^2$$

Then the total variance is

$$T = \bar{S} + (1 + m^{-1})B$$

SUMMARY

- FCS provides a flexible method for dealing with missing values
- Other methods exist:

  Full Information Likelihood Strategy: integrates out the missing data when fitting the desired model; Needs a specification of the full data likelihood; Works under MAR and MNAR (with more assumptions)

  Inverse probability weighting: Weight cases by their probability of having complete data

  E.g., an individual with a low probability of being a complete case will receive more weight (and count, say, 3 times)

  You need a model for this probability