Universiteit Utrecht

Imputation of missing data under missing not at random assumption & sensitivity analysis

S. Jolani

Department of Methodology and Statistics, Utrecht University, the Netherlands

Advanced Multiple Imputation, Utrecht, May 2013

Outline

Why missing not at random (MNAR)?

- There might be a reason to believe that responders differ from non-responders, even after accounting for the observed information
- Some examples:

- Income - some people may not reveal their salaries
- Blood pressure - the blood pressure is measured less frequently for patients with lower blood pressures
- Depression - some patients might dropout because they believe the treatment is not effective

Notation

Y: incomplete variable

R: response indicator ($R = 1$ if Y is observed)

X: fully observed covariate

Y_{obs} and Y_{mis}: the observed and missing parts of Y

A general strategy

Y and R must be modeled **jointly** (Rubin, 1976) under an MNAR assumption

so

$$P(Y, R)$$

Why the classical MI does not work?

Imputation under MAR

$$P(Y|X, R = 0) = P(Y|X, R = 1)$$

Why the classical MI does not work?

Imputation under MAR

$$P(Y|X, R = 0) = P(Y|X, R = 1)$$

Imputation under MNAR

$$P(Y|X, R = 0) \neq P(Y|X, R = 1)$$

Models for nonignorable nonresponse

Two general approaches (there are some more):

1. Selection models (Heckman, 1976)
2. Pattern mixture-models (Rubin, 1977)

Selection model

$$P(Y, R; \xi, \omega) = P(Y; \xi)P(R|Y; \omega),$$

where the parameters ξ and ω are *a priori* independent.

$P(Y; \xi)$	distribution for the full data
$P(R\|Y; \omega)$	response mechanism (selection function)

Selection model

Imputation model under MNAR

$$P(Y_{mis}|X, Y_{obs}, R)$$

where

$$P(Y_{mis}|X, Y_{obs}, R) = \frac{P(Y_{mis}|X, Y_{obs})P(R|X, Y)}{\int P(Y_{mis}|X, Y_{obs})P(R|X, Y)\partial Y_{mis}}$$

Selection model

Imputation model under MNAR

$$P(Y_{mis}|X, Y_{obs}, R)$$

A simple but possibly inefficient approach (Rubin, 1987):

1. Draw a candidate $Y_i^* \sim P(Y_i|X_i; \xi = \xi^*)$
2. Calculate $p_i^* = P(R_i = 1|X_i, Y_i = Y_i^*; \omega)$
3. Draw $R_i^* \sim Ber(1, p_i^*)$
4. Impute Y_i^* if $R_i^* = 0$ otherwise return to (1)

Pattern mixture model

$$P(Y, R; \psi, \theta) = P(R; \psi)P(Y|R; \theta),$$

where the parameters ψ and θ are *a priori* independent.

$P(Y\|X, R = 1; \theta_1)$	distribution for the observed data
$P(Y\|X, R = 0; \theta_0)$	distribution for the missing data
$P(R; \psi)$	marginal response probability

Pattern mixture model

The general procedure (Rubin, 1977):

1. Draw θ_1^* from its posterior distribution using $P(Y|X, R = 1; \theta_1)$
2. Specify the posterior $P(\theta_0|\theta_1)$ *a priori* (e.g., $\theta_0 = \theta_1 + k$ where k is a fixed constant)
3. Draw $\theta_0^* \sim P(\theta_0|\theta_1^*)$
4. Impute Y_{mis} from $P(Y|X, R = 0; \theta_0^*)$

An example

Suppose Y is an incomplete variable (continuous)

$$Y_{obs} \sim N(\mu_1, \sigma_1^2), \quad Y_{mis} \sim N(\mu_0, \sigma_0^2)$$

where $\theta_1 = (\mu_1, \sigma_1^2)$ and $\theta_0 = (\mu_0, \sigma_0^2)$. Now, if we define

$$\mu_0 = \mu_1 + k_1, \quad \sigma_0^2 = k_2\sigma_1^2$$

where k_1 and k_2 are fixed and known values (sensitivity parameters).

An example

Suppose Y is an incomplete variable (continuous)

$$Y_{obs} \sim N(\mu_1, \sigma_1^2), \quad Y_{mis} \sim N(\mu_0, \sigma_0^2)$$

where $\theta_1 = (\mu_1, \sigma_1^2)$ and $\theta_0 = (\mu_0, \sigma_0^2)$. Now, if we define

$$\mu_0 = \mu_1 + k_1, \quad \sigma_0^2 = k_2 \sigma_1^2$$

where k_1 and k_2 are fixed and known values (sensitivity parameters).

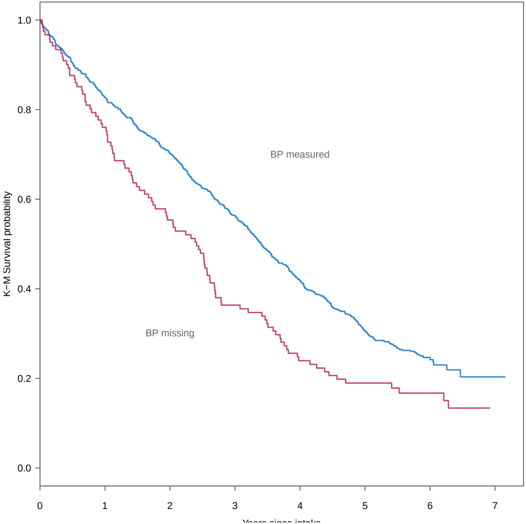
Sensitivity analysis:
repeat the analysis for different choices of k_1 and k_2

Application: Leiden 85+

- Leiden 85+ cohort study
- N=1236, 85+ on Dec. 1, 1986
- N=956 were visited (1987-1989)
- Blood pressure (BP) is missing for 121 patients

* Do anti-hypertensive drugs shorten life in the oldest old?
* Scientific interest: Mortality risk as function of BP and age

Survival probability by response group



Source: van Buuren et al. (1999)

Leiden 85+

From the data we see

- Those with no BP measured die earlier
- Those that die early and that have no hypertension history have fewer BP measurements

Thus, imputations of BP under MAR could be too high values.
We need to lower the imputed values of BP, and study the influence on the outcome

A simple model to shift imputations

Y: BP
X: age, hypertension, haemoglobin, and etc

Specify $P(Y|X,R)$

Model		
1	$Y = X\beta + \epsilon$	$R = 1$
2	$Y = X\beta + \delta + \epsilon$	$R = 0$

Combined formulation:

$$Y = X\beta + (1 - R)\delta + \epsilon$$

δ cannot be estimated (sensitivity parameter)

Numerical example

Table IV. Numerical example of an NMAR non-response mechanism, when there are more missing data for lower blood pressures

Y	Selection model		Mixture model	
Class midpoint of Systolic BP (mmHg)	$p(R=0\|BP)$	$p(BP)$	$p(BP\|R=1)$	$p(BP\|R=0)$
100	0·35	0·02	0·01	0·06
110	0·30	0·03	0·02	0·07
120	0·25	0·05	0·04	0·10
130	0·20	0·10	0·09	0·16
140	0·15	0·15	0·15	0·19
150	0·10	0·30	0·31	0·25
160	0·08	0·15	0·16	0·10
170	0·06	0·10	0·11	0·05
180	0·04	0·05	0·05	0·02
190	0·02	0·03	0·03	0·00
200	0·00	0·02	0·02	0·00
Mean (mmHg)		150	151·6	138·6

Source: van Buuren et al. (1999)

How to impute under MNAR in MICE?


```
> delta <- c(0,-5,-10,-15,-20)
> post <- mice(leiden85,maxit=0)$post
> imp.all <- vector("list", length(delta))
> for (i in 1:length(delta)) {
+   d <- delta[i]
+   cmd <- paste("imp[[j]][,i] <- imp[[j]][,i] +",d)
+   post["bp"] <- cmd
+   imp <- mice(leidan85, post=post, seed=i*22, print=FALSE)
+   imp.all[[i]] <- imp
+ }
```

Leiden 85+: Sensitivity analysis

Table V. Mean and standard deviation of the observed and imputed blood pressures. The statistics of imputed BP are pooled over $m = 5$ multiple imputations

	N	δ	SBP Mean	SBP SD	DBP Mean	DBP SD
Observed BP	835		152·9	25·7	82·8	13·1
Imputed BP	121	0	151·1	26·2	81·5	14·0
	121	−5	142·3	24·6	78·4	13·7
	121	−10	135·9	24·7	78·2	12·8
	121	−15	128·6	25·0	75·3	12·9
	121	−20	122·3	25·2	74·0	12·1

Source: van Buuren et al. (1999)

Drawn indicator imputation

Combined formulation: $Y = X\beta + (1 - R)\delta + \epsilon$

if $\epsilon \sim N(0, \sigma^2)$, then

$$Y_{obs} \sim N(X\beta, \sigma^2) \quad (1)$$
$$Y_{mis} \sim N(X\beta + \delta, \sigma^2) \quad (2)$$

$$\begin{aligned} \text{logit}\{P(R = 1|X, Y)\} &= \log[\frac{P(R = 1)P(Y|X, R = 1)}{P(R = 0)P(Y|X, R = 0)}] \\ &= \psi_0 + \psi_1 Y + \psi_2 X, \quad (3) \end{aligned}$$

where $\psi_1 = \delta/\sigma^2$ so that $\delta = \psi_1 * \sigma^2$.

Drawn indicator imputation

Assume $P(R = 1|X, Y)$ is known (unrealistic!)

Y	R	1 - P(R = 1|X, Y)	R_1
200	1	.00	1
195	1	.02	1
183	1	.06	1
180	1	.09	1
176	1	.10	0
160	1	.15	0
140	1	.20	0
.	0	.25	1
.	0	.30	1
.	0	.38	0
.	0	.42	0
.	0	.45	0
.	0	.50	0

Drawn indicator imputation

Gr	Y	R	R_1	E(Y|X,R, R_1)
1	200	1	1	μ_{11}
	195	1	1	
	183	1	1	
	180	1	1	
2	176	1	0	μ_{10}
	160	1	0	
	140	1	0	
3	.	0	1	μ_{01}
	.	0	1	
4	.	0	0	μ_{00}
	.	0	0	
	.	0	0	
	.	0	0	

It can be shown that

$$\begin{aligned} \mu_{10} &= \mu_{01} \\ \mu_{11} - \mu_{10} &\simeq \mu_{01} - \mu_{00} \end{aligned}$$

The idea?

- Impute group 3 from group 2
- Impute group 4 from groups 2 and 1

Drawn indicator imputation

But in reality $P(R = 1|X, Y)$ is unknown

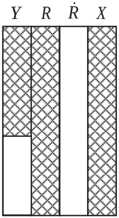


Figure : The schematic representation of the data

Drawn indicator imputation

But in reality $P(R = 1|X, Y)$ is unknown

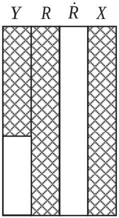


Figure : The schematic representation of the data

Fully Conditional Specification:

$$\begin{aligned} Y &\sim P(Y|X, R, R_1) \\ R_1 &\sim P(R_1|X, Y) \end{aligned}$$

Drawn indicator imputation

1. Impute initially missing values (Y^*)
2. Draw $\dot{R}$ from a Bernouli process ($\dot{R} \sim Ber(1, \pi)$) where $\pi = P(R = 1|X, Y^*)$
3. Using groups 1 and 2, estimate β and δ from $E(Y|X, R = 1, R_1 = r_1) = X\beta + \delta(r_1 - 1), \quad r_1 = 0, 1$
4. Draw $\dot{\beta}$ from its posterior distribution for a given prior for β
5. Predict the missing data for group 3 using $X\dot{\beta} - \hat{\delta}$
6. Predict the missing data for group 4 using $X\dot{\beta} - 2\hat{\delta}$
7. Impute the missing data by adding an appropriate amount of noise to the predicted values
8. Return to Step 2

How to implement the drawn indicator method in MICE?


```
> mice(data, meth = "ri")
```

The RI function:

```
> mice.impute.ri(y, ry, x, ri.maxit = 10, ...)
```

Note:

1. only for continuous variables (the current version)
2. the same covariates for both models (the current version)

Leiden 85+

- Summary

  Participants: 956
  Observed BP: 835
  Missing BP: 121
- Imputation model

  BP $\sim$ sex, age, hypertension, haemoglobin, etc.
- Missingness mechanism

  logit$\{P(R = 1|Y, X)\} \sim$ BP, type of residence, ADL, previous hypertension, etc.

- Number of iterations: 10
- Number of multiple imputations: 50

Leiden 85+

Table : Mean and standard error (SE) for the systolic blood pressure using CC, MI and RI

	Total		Imputed	
Method	Mean	SE	Mean	SE
CC	152.893	0.892	-	-
MI	152.473	0.924	149.47	2.409
RI	151.075	1.109	139.06	2.438

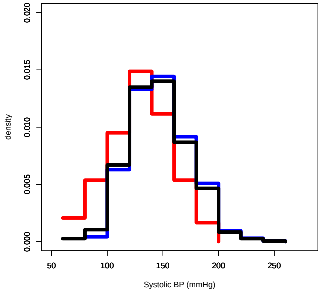
Leiden 85+

160	0·08	0·15	0·16	0·10
170	0·06	0·10	0·11	0·05
180	0·04	0·05	0·05	0·02
190	0·02	0·03	0·03	0·00
200	0·00	0·02	0·02	0·00
Mean (mmHg)		150	151·6	138·6

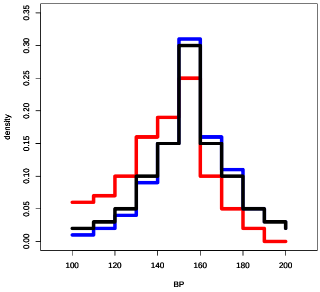
An interesting result:

Using the RI method, we are able to estimate $\hat{\delta} = 139.1 - 152.9 = -13.8$. This value is very similar to the amount of the adjustment in van Buuren et al. (1999) based on a numerical example.

Effect of response mechanism on BP



Leiden85+ (the drawn indicator method)



Numerical example (van Buuren et al. 1999)

A summary of the models under MNAR

1. All methods for the incomplete data under MNAR make unverified assumptions
2. Selection model: the distribution of the full data
3. Pattern mixture: the distribution of the missing data
4. Drawn indicator: the distribution of the selection function

General advice on MNAR

1. Why is the ignorability assumption is suspected? (why MNAR assumption)
2. Include as much data as possible in the imputation model
3. Limit the possible non-ignorable alternatives