STAT 403

May 9, 2025

Lecture 15

Imputation

HW5 – TB Posted

Project TB Posted

Lecture Notes VII – Missing Data and Imputation

Marina Meila
mmp@stat.washington.edu

Department of Statistics
University of Washington

April 2025

The missing data problem

MCAR ✓

MAR ✓

MNAR ? ✗

Imputation for MAR $p(y \mid x, R=1)$

Imputation for monotone missingness (MAR)

What is the distribution of R?

- Let data $\mathcal{D} = \{(x^1, y^1), \dots (x^n, y^n)\} \cup \{x_{n+1}, \dots x_{n+m}\}$
  - complete data$\{(x^1, y^1), \dots (x^n, y^n)\}$ with $R^{1:n} = 1$
  - data with missing values $\{x_{n+1}, \dots x_{n+m}\}$ with $R^{n+1:n+m} = 0$

MCAR Missing Completely At Random $R \perp (X, Y)$

- no bias if only observed Y values $\{y^1, \dots y^n)\}$ used
- Can also use imputation as in MAR case

MAR Missing At Random $R \perp Y \,|\, X$

- **bias possible** if only observed Y values $\{y^1, \dots y^n)\}$ are used because $R \not\perp Y$
- $\{(x^1, y^1, R^1), \dots (x^n, y^n, R^n)\}$ together with $\{(x^{n+1}, R^{n+1}), \dots (x^{n+m}, R^{n+m})\}$ has information about $\{y^{n+1}, \dots y^{n+m}\}$
- Probability distribution of missing values $Y^{n+1:n+m}$ is

$$Y \sim p(Y \,|\, X, R = 0) \quad \text{and} \quad \boxed{p(Y \,|\, X, R = 0) \; = \; p(Y \,|\, X, R = 1)} \tag{1}$$

Imputation for MAR with Monothone Missing data

- Example Clinical trials with drop-out Each patient i is observed for a $t = 1, 2, \ldots T^i$ with $T^i \leq p$. Hence $y^i = (y^i_1, \ldots y^i_{T^i})$.
- T can depend on the previous values $y^{1:T}$ but not on the future values $y^{T=1:p}$.

- T^i is the missingness variable
- $x^i_{1:t}$ is the "always observed" data
- $x^i_{t+1:p}$ is the missing data

Setting ("clinical trial") longitudinal $x \in \mathbb{R}^p$

variables $[x_1, x_2, \ldots x_t \,|\, \ldots x_p]$ dropped

data $x^i_1 \ldots\ldots x^i_t \,|\, \underbrace{y^i_{t+1} \cdots y^i_p}_{\text{missing}}$ t^i

missingness $\underbrace{R^i_1 \ldots\ldots R^i_t}_{1}$ $\underbrace{R^i_{t+1} \cdots R^i_p}_{=0}$ $\Rightarrow t^i$

MAR $\quad Y \perp R \,|\, X$ $\qquad p(y \,|\, X, R=1)$ for obs data

$x_{t+1:p} \equiv y_{t+1:p} \perp t \,|\, x_{1:t}$ $\qquad t = 1, 2, 3, \ldots p-1$ levels of missingness

Imputation for MAR with Monothone Missing data

- Example Clinical trials with drop-out Each patient i is observed for a $t = 1, 2, \dots T^i$ with $T^i \le p$. Hence $y^i = (y_1^i, \dots y_{T^i}^i)$.
- T can depend on the previous values $y^{1:T}$ but not on the future values $y^{T=1:p}$.

- T^i is the missingness variable
- $x_{1:t}^i$ is the "always observed" data
- $x_{t+1:p}^i$ is the missing data

$\boxed{T = \text{stopping time}}$

for $t = 1, 2, \dots p-1$

want: estimate $p(x_{t+1} \mid x_{1:t}, T=t) = p(x_{t+1} \mid x_{1:t}, T>t)$

missing

want

obs

Alg 1

i=1 0 → $t^1 = 1$

i=2 0 1 0 $t^2 = 3$

i=3 1 1

0 1 1 1 0 0 $t^4 = 6$

0 0 1 1 0 5

0 1 x_3^6 2

x_1^6, x_2^6 missing

$$p(x_3^6 \mid x_1^6, x_2^6, T=2) = \left(\frac{1}{2}, \frac{1}{2}\right)$$

$$p(x_2 \mid x_1, T=1) = \begin{cases} x_1 = 1 & ? \\ x_1 = 0 & \left(\frac{3}{4}, \frac{1}{4}\right) \end{cases}$$

1

0

010 $\cancel{X}_4^2$

1

011100

00110

01

$p(X_4 \mid X_{1:3}, T=3) = ?$ no data !!

010

no data either

Need additional assumptions

Ex. 1 $\quad X_{t+1} \mid \#\{x_j = 0,\ j = 1:t\}$

Ex. 2. $\quad X_{t+1} \mid X_t$

Ex 3. $\quad X_{t+1} \mid X_t$ same way for all t

$$p(x_{t+1} \mid x_t) \equiv p(x_{t+1} \mid x_t, T) = \begin{pmatrix} \frac{4}{6} & \frac{2}{6} \end{pmatrix}$$

$\qquad\qquad 1 \quad 0$

use this Ex1

Back to original question

θ = param of interest

Eg $\quad \theta = E[X_p]$

$\theta = P[\text{longest sequence of 1's}]$

...

Need $\quad P[X] \equiv P[X_{1:p}]$

$X = X_{1:t} \mid \underbrace{X_{t+1:p}}_{\text{missing}}$

t_s

can be imputed

$$P(X) = P(X_{1:t})\,P(X_{t+1} \mid X_{1:t})\,P(X_{t+2} \mid X_{t+1}, X_{1:t}) \cdots P(X_p \mid X_{1:p-1})$$

(chain rule)

known $\leftarrow$ Alg 1

Need $P[X] = P[X_{1:p}]$

$$x = x_{1:t} \,\Big|_{t}\, \underbrace{x_{t+1:p}}_{\text{missing}}$$

$$P(x) = P(x_{1:t})\,P(x_{t+1} \mid x_{1:t})\,P(x_{t+2} \mid x_{t+1}, x_{1:t}) \cdots P(x_p \mid x_{1:p-1})$$

Moreover T also has info about $x_{1:T}$, therefore

$$P(x) = \sum_{t=1}^{p} p(x, t) =$$

$$= \sum_{t=1}^{p} P(T=t)\,P(x \mid T=t)$$

$$= \sum_{t=1}^{p} \underbrace{P(T=t)}_{\text{from data}}\ \underbrace{P(x_{1:t} \mid T=t)}_{\text{from data: sequences of length } t}\ \underbrace{P(x_{t+1:p} \mid x_{1:t})}_{\text{by Alg 1}}$$

$T=1$ | 1 0

0 1 0

| 1

0 1 1 1 0 0

0 0 1 1 0

|| 0 1

T	1	2	3	4	5	6
$P(T)$:	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$		$\frac{1}{6}$	$\frac{1}{6}$

$P(X_1 \mid T=1)$	0	1
	$\frac{1}{2}$	$\frac{1}{2}$

$$P(X_1 X_2 \mid T=2) = \begin{cases} 1 & x_1 x_2 = 01 \\ 0 & \text{otherwise} \end{cases}$$

etc

Finally: → use $P(x)$ to estimate θ

↓ use imputation ⇒ complete sample ⇒ estimate θ