Statistical Methods for Analysis with Missing Data

Lecture 2: general setup, notation, missing-data mechanisms

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY of WASHINGTON

Previous Lecture

$$\underbrace{p(y)}_{\text{what we want}} = p(y \mid R=0)p(R=0) + \underbrace{p(y \mid R=1)}_{\text{what we can get}}p(R=1)$$

We cannot recover $p(y \mid R=0)$ nor $p(y)$ from observed data alone

The fundamental problem of inference with missing data: it is impossible without extra, usually untestable, assumptions on how missingness arises

Today's Lecture

- General setup, notation
- Missing-data mechanisms

Reading: pages 14 – 22, Ch. 1, of Davidian and Tsiatis

Outline

Study Variables and Response Indicators

Gender	Age	Income	R_{Gender}	R_{Age}	R_{Income}	...
F	25	60,000	1	1	1	...
M	?	?	1	0	0	...
?	51	?	0	1	0	...
F	?	150,300	1	0	1	...
...	...	...	...	...	...	...

Study Variables and Response Indicators

- $Z = (Z_1, \ldots, Z_K)$: the *study variables*, or the variables that we intend to measure on each individual
  - Each Z_k, $k = 1, \ldots, K$, is a *block of variables* that are jointly missing/observed
- $R = (R_1, \ldots, R_K)$: the *response indicators*
  - Each R_k, $k = 1, \ldots, K$, is an indicator of whether Z_k is observed

$$R_k = \begin{cases} 1 & \text{if } Z_k \text{ is observed,} \\ 0 & \text{if } Z_k \text{ is missing.} \end{cases}$$

Study Variables and Response Indicators

- $Z = (Z_1, \ldots, Z_K)$: the *study variables*, or the variables that we intend to measure on each individual
  - Each Z_k, $k = 1, \ldots, K$, is a *block of variables* that are jointly missing/observed
- $R = (R_1, \ldots, R_K)$: the *response indicators*
  - Each R_k, $k = 1, \ldots, K$, is an indicator of whether Z_k is observed

$$R_k = \begin{cases} 1 & \text{if } Z_k \text{ is observed,} \\ 0 & \text{if } Z_k \text{ is missing.} \end{cases}$$

Sample Data

Gender	Age	Income	R_{Gender}	R_{Age}	R_{Income}	...
F	25	60,000	1	1	1	...
M	?	?	1	0	0	...
?	51	?	0	1	0	...
F	?	150,300	1	0	1	...
...	...	...	...	...	...	...

For each individual $i = 1, \dots, n$, we define

- Study variables: $Z_i = (Z_{i1}, \dots, Z_{iK})$
- Response indicators: $R_i = (R_{i1}, \dots, R_{iK})$

Sample Data

- We assume the *full sample* are independent and identically distributed (i.i.d.) draws

$$\{(Z_i, R_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} F$$

  from some distribution F

- Of course, this an idealized scenario: we typically cannot fully observe Z_i

- In this lecture, we delete the subindex i to talk about a generic draw from F

Response and Missingness Patterns

- Each of the components of Z can either be missing or observed, so in general

$$R = (R_1, \ldots, R_K) \in \{0,1\}^K$$

Example: if $K = 2$, $\{0,1\}^2 = \{(0,0),(1,0),(0,1),(1,1)\}$

- $r = (r_1, \ldots, r_K)$: generic element of $\{0,1\}^K$, a *response pattern*
  - Sometimes we write r as a string $r = r_1 \ldots r_K$
  - e.g., $r = (0,1,0) \equiv 010$
- $\bar{R} = (1 - R_1, \ldots, 1 - R_K)$: the *missingness indicators*
- $\bar{r}$: generic value of $\bar{R}$, a *missingness pattern*

Response and Missingness Patterns

- Each of the components of Z can either be missing or observed, so in general

$$R = (R_1, \ldots, R_K) \in \{0,1\}^K$$

Example: if $K = 2$, $\{0,1\}^2 = \{(0,0),(1,0),(0,1),(1,1)\}$

- $r = (r_1, \ldots, r_K)$: generic element of $\{0,1\}^K$, a *response pattern*
  - Sometimes we write r as a string $r = r_1 \ldots r_K$
  - e.g., $r = (0,1,0) \equiv 010$
- $\bar{R} = (1 - R_1, \ldots, 1 - R_K)$: the *missingness indicators*
- $\bar{r}$: generic value of $\bar{R}$, a *missingness pattern*

Response and Missingness Patterns

- Each of the components of Z can either be missing or observed, so in general
$$R = (R_1, \ldots, R_K) \in \{0,1\}^K$$

Example: if $K = 2$, $\{0,1\}^2 = \{(0,0),(1,0),(0,1),(1,1)\}$

- $r = (r_1, \ldots, r_K)$: generic element of $\{0,1\}^K$, a *response pattern*
  - Sometimes we write r as a string $r = r_1 \ldots r_K$
  - e.g., $r = (0,1,0) \equiv 010$
- $\bar{R} = (1 - R_1, \ldots, 1 - R_K)$: the *missingness indicators*
- $\bar{r}$: generic value of $\bar{R}$, a *missingness pattern*

Response and Missingness Patterns

- Each of the components of Z can either be missing or observed, so in general
$$R = (R_1, \ldots, R_K) \in \{0,1\}^K$$

Example: if $K = 2$, $\{0,1\}^2 = \{(0,0),(1,0),(0,1),(1,1)\}$

- $r = (r_1, \ldots, r_K)$: generic element of $\{0,1\}^K$, a *response pattern*
  - Sometimes we write r as a string $r = r_1 \ldots r_K$
  - e.g., $r = (0,1,0) \equiv 010$
- $\bar{R} = (1 - R_1, \ldots, 1 - R_K)$: the *missingness indicators*
- $\bar{r}$: generic value of $\bar{R}$, a *missingness pattern*

Response and Missingness Patterns

- Each of the components of Z can either be missing or observed, so in general
$$R = (R_1, \ldots, R_K) \in \{0,1\}^K$$

Example: if $K = 2$, $\{0,1\}^2 = \{(0,0),(1,0),(0,1),(1,1)\}$

- $r = (r_1, \ldots, r_K)$: generic element of $\{0,1\}^K$, a *response pattern*
  - Sometimes we write r as a string $r = r_1 \ldots r_K$
  - e.g., $r = (0,1,0) \equiv 010$
- $\bar{R} = (1 - R_1, \ldots, 1 - R_K)$: the *missingness indicators*
- $\bar{r}$: generic value of $\bar{R}$, a *missingness pattern*

Notation Example: Regression

Say

$$Z = (Y, X) = (Y, X_1, \ldots, X_p)$$

where Y is a response, and X are covariates

- Say only the outcome Y can be missing, then
  - $Z = (Z_1, Z_2), \quad Z_1 = Y, \quad Z_2 = X$
  - $R = (R_1, R_2) \in \{(0,1), (1,1)\}$
  - Alternatively, we could define $R \in \{0, 1\}$, $R = 1$ if Y is observed

- Say outcome Y and covariates X can be missing (all covariates at the same time), then
  - $Z = (Z_1, Z_2), \quad Z_1 = Y, \quad Z_2 = X$
  - $R = (R_1, R_2) \in \{0, 1\}^2$

- Say outcome Y and individual covariates $X_1, \ldots, X_p$ can be missing (regardless of the missing status of others), then
  - $Z = (Z_1, Z_2, \ldots, Z_{p+1}), \quad Z_1 = Y, \quad Z_2 = X_1, \ \ldots, \ Z_{p+1} = X_p$
  - $R = (R_1, \ldots, R_{p+1}) \in \{0, 1\}^{p+1}$

Notation Example: Regression

Say

$$Z = (Y, X) = (Y, X_1, \ldots, X_p)$$

where Y is a response, and X are covariates

- Say only the outcome Y can be missing, then
  - $Z = (Z_1, Z_2), \quad Z_1 = Y, \quad Z_2 = X$
  - $R = (R_1, R_2) \in \{(0,1), (1,1)\}$
  - Alternatively, we could define $R \in \{0,1\}$, $R = 1$ if Y is observed
- Say outcome Y and covariates X can be missing (all covariates at the same time), then
  - $Z = (Z_1, Z_2), \quad Z_1 = Y, \quad Z_2 = X$
  - $R = (R_1, R_2) \in \{0,1\}^2$
- Say outcome Y and individual covariates $X_1, \ldots, X_p$ can be missing (regardless of the missing status of others), then
  - $Z = (Z_1, Z_2, \ldots, Z_{p+1}), \quad Z_1 = Y, \quad Z_2 = X_1, \quad \ldots, \quad Z_{p+1} = X_p$
  - $R = (R_1, \ldots, R_{p+1}) \in \{0,1\}^{p+1}$

Notation Example: Regression

Say

$$Z = (Y, X) = (Y, X_1, \dots, X_p)$$

where Y is a response, and X are covariates

- Say only the outcome Y can be missing, then
  - $Z = (Z_1, Z_2), \quad Z_1 = Y, \quad Z_2 = X$
  - $R = (R_1, R_2) \in \{(0,1), (1,1)\}$
  - Alternatively, we could define $R \in \{0, 1\}$, $R = 1$ if Y is observed
- Say outcome Y and covariates X can be missing (all covariates at the same time), then
  - $Z = (Z_1, Z_2), \quad Z_1 = Y, \quad Z_2 = X$
  - $R = (R_1, R_2) \in \{0, 1\}^2$
- Say outcome Y and individual covariates $X_1, \dots, X_p$ can be missing (regardless of the missing status of others), then
  - $Z = (Z_1, Z_2, \dots, Z_{p+1}), \quad Z_1 = Y, \quad Z_2 = X_1, \quad \dots, \quad Z_{p+1} = X_p$
  - $R = (R_1, \dots, R_{p+1}) \in \{0, 1\}^{p+1}$

Notation Example: Longitudinal Study

Study participants' characteristics are to be measured at T times

- Z_j: measurements taken at time t_j
- R_j: indicator of whether participant shows up at time t_j
- If missingness only comes from subjects dropping out
  - Drop out at time t_j: $Z_1, \ldots, Z_{j-1}$ observed; $Z_j, \ldots, Z_T$ not observed
  - $R = (R_1, \ldots, R_T) \in \{(1, 0, \ldots, 0), (1, 1, 0, \ldots, 0), \ldots, (1, 1, \ldots, 1)\}$
  - Can be uniquely summarized by the drop out time $D = 1 + \sum_{j=1}^{T} R_j$
- If participants sporadically show up
  - $R = (R_1, \ldots, R_T) \in \{0, 1\}^T$

Notation Example: Longitudinal Study

Study participants' characteristics are to be measured at T times

- Z_j: measurements taken at time t_j
- R_j: indicator of whether participant shows up at time t_j
- If missingness only comes from subjects dropping out
  - Drop out at time t_j: $Z_1, \ldots, Z_{j-1}$ observed; $Z_j, \ldots, Z_T$ not observed
  - $R = (R_1, \ldots, R_T) \in \{(1,0,\ldots,0), (1,1,0,\ldots,0), \ldots, (1,1,\ldots,1)\}$
  - Can be uniquely summarized by the drop out time $D = 1 + \sum_{j=1}^{T} R_j$
- If participants sporadically show up
  - $R = (R_1, \ldots, R_T) \in \{0,1\}^T$

Notation Example: Longitudinal Study

Study participants' characteristics are to be measured at T times

- Z_j: measurements taken at time t_j
- R_j: indicator of whether participant shows up at time t_j
- If missingness only comes from subjects dropping out
  - Drop out at time t_j: $Z_1, \ldots, Z_{j-1}$ observed; $Z_j, \ldots, Z_T$ not observed
  - $R = (R_1, \ldots, R_T) \in \{(1, 0, \ldots, 0), (1, 1, 0, \ldots, 0), \ldots, (1, 1, \ldots, 1)\}$
  - Can be uniquely summarized by the drop out time $D = 1 + \sum_{j=1}^{T} R_j$
- If participants sporadically show up
  - $R = (R_1, \ldots, R_T) \in \{0, 1\}^T$

Missing and Observed Data

Given $R = r$

- $Z_{(r)}$: observed values
- $Z_{(\bar{r})}$: missing values

Example:

- $Z = (Z_1, Z_2, Z_3)$
- If $r = 010$, $\quad Z_{(r)} = Z_{(010)} = Z_2$, $\quad$ and $\quad Z_{(\bar{r})} = Z_{(101)} = (Z_1, Z_3)$

HW1: write down $Z_{(r)}$ and $Z_{(\bar{r})}$ for all possible values of $r \in \{0, 1\}^3$

Missing and Observed Data

Given $R = r$

- $Z_{(r)}$: observed values
- $Z_{(\bar{r})}$: missing values

Example:

- $Z = (Z_1, Z_2, Z_3)$
- If $r = 010$, $\quad Z_{(r)} = Z_{(010)} = Z_2, \quad$ and $\quad Z_{(\bar{r})} = Z_{(101)} = (Z_1, Z_3)$

HW1: write down $Z_{(r)}$ and $Z_{(\bar{r})}$ for all possible values of $r \in \{0, 1\}^3$

Missing and Observed Data

Given $R = r$

- $Z_{(r)}$: observed values
- $Z_{(\bar{r})}$: missing values

Example:

- $Z = (Z_1, Z_2, Z_3)$
- If $r = 010$, $\quad Z_{(r)} = Z_{(010)} = Z_2, \quad$ and $\quad Z_{(\bar{r})} = Z_{(101)} = (Z_1, Z_3)$

HW1: write down $Z_{(r)}$ and $Z_{(\bar{r})}$ for all possible values of $r \in \{0, 1\}^3$

Observed Data

Given that R is random, the *observed data* are obtained as realizations of

$$(Z_{(R)}, R)$$

We can think of the generative process

$$Z \implies R \implies (Z_{(R)}, R)$$

- HW1: explain what is the difference between $(Z_{(R)}, R)$ and $(Z_{(r)}, R = r)$ for a fixed value r
- HW1:
  a) say $Z = (Z_1, Z_2)$, $Z_1 \in \{1, 2\}$, $Z_2 \in \{A, B\}$, $R \in \{0, 1\}^2$. Write down all the elements of the sample space of $(Z_{(R)}, R)$.
  b) Describe the sample space of $(Z_{(R)}, R)$ if instead $Z \in \mathbb{R}^2$.

Observed Data

Given that R is random, the *observed data* are obtained as realizations of

$$(Z_{(R)}, R)$$

We can think of the generative process

$$Z \implies R \implies (Z_{(R)}, R)$$

- HW1: explain what is the difference between $(Z_{(R)}, R)$ and $(Z_{(r)}, R = r)$ for a fixed value r
- HW1:
  a) say $Z = (Z_1, Z_2)$, $Z_1 \in \{1, 2\}$, $Z_2 \in \{A, B\}$, $R \in \{0, 1\}^2$. Write down all the elements of the sample space of $(Z_{(R)}, R)$.
  b) Describe the sample space of $(Z_{(R)}, R)$ if instead $Z \in \mathbb{R}^2$.

Observed Data

Given that R is random, the *observed data* are obtained as realizations of

$$(Z_{(R)}, R)$$

We can think of the generative process

$$Z \implies R \implies (Z_{(R)}, R)$$

- HW1: explain what is the difference between $(Z_{(R)}, R)$ and $(Z_{(r)}, R = r)$ for a fixed value r
- HW1:
  a) say $Z = (Z_1, Z_2)$, $Z_1 \in \{1, 2\}$, $Z_2 \in \{A, B\}$, $R \in \{0, 1\}^2$. Write down all the elements of the sample space of $(Z_{(R)}, R)$.
  b) Describe the sample space of $(Z_{(R)}, R)$ if instead $Z \in \mathbb{R}^2$.

Observed Data

Given that R is random, the *observed data* are obtained as realizations of

$$(Z_{(R)}, R)$$

We can think of the generative process

$$Z \implies R \implies (Z_{(R)}, R)$$

- HW1: explain what is the difference between $(Z_{(R)}, R)$ and $(Z_{(r)}, R = r)$ for a fixed value r
- HW1:
  a) say $Z = (Z_1, Z_2)$, $Z_1 \in \{1, 2\}$, $Z_2 \in \{A, B\}$, $R \in \{0, 1\}^2$. Write down all the elements of the sample space of $(Z_{(R)}, R)$.

  b) Describe the sample space of $(Z_{(R)}, R)$ if instead $Z \in \mathbb{R}^2$.

The (Z_{obs}, Z_{mis}) Notation

- To formally characterize the observed data we need to use the response vector R
- Yet, a large portion of the literature on missing data define the observed and missing data as

$$Z = (Z_{obs}, Z_{mis})$$

  - Z_{obs}: observed values, so "$Z_{obs} = Z_{(R)}$"
  - Z_{mis}: missing values, so "$Z_{mis} = Z_{(\bar{R})}$"
- This notation is convenient for its simplicity, but in this course we avoid it, as Z_{obs} and Z_{mis} do not explicitly indicate how they relate to R

Notation Example: Longitudinal Study

If missingness comes only from subjects dropping out

- Missingness patterns are uniquely summarized by the drop out time $D = 1 + \sum_{j=1}^{T} R_j$
- The *observed data* are obtained as realizations of

$$(Z_{(D)}, D)$$

where, if $D = d$, $Z_{(d)} = (Z_1, \ldots, Z_{d-1})$

Distributions of Interest

- Full-data distribution: joint distribution of (Z, R)
  - Density: $p(z, r) = p(z \mid r)p(r) = p(r \mid z)p(z)$
- Davidian and Tsiatis refer to the distribution of Z as the full-data distribution, but R is also data!
- Missing-data mechanism or missingness mechanism: conditional distribution of $R \mid Z$
  - Density: $p(r \mid z)$

Distributions of Interest

- Full-data distribution: joint distribution of (Z, R)
  - Density: $p(z, r) = p(z \mid r)p(r) = p(r \mid z)p(z)$
- Davidian and Tsiatis refer to the distribution of Z as the full-data distribution, but R is also data!
- Missing-data mechanism or missingness mechanism: conditional distribution of $R \mid Z$
  - Density: $p(r \mid z)$

Distributions of Interest

- Full-data distribution: joint distribution of (Z, R)
  - Density: $p(z, r) = p(z \mid r)p(r) = p(r \mid z)p(z)$
- Davidian and Tsiatis refer to the distribution of Z as the full-data distribution, but R is also data!
- Missing-data mechanism or missingness mechanism: conditional distribution of $R \mid Z$
  - Density: $p(r \mid z)$

Notation for Density Functions

For simplicity we use $p(\cdot)$ for technically different functions

- $p(z) \equiv p_Z(z)$
- $p(z, r) \equiv p_{Z,R}(z, r)$
- $p(r \mid z) \equiv p_{R|Z}(r \mid z)$

Interpretations should be clear from the arguments passed to them

Outline

Notation

Missing-Data Mechanisms

Missing-Data Mechanisms: A Bit of History

- Missing data was largely seen as a computational issue: “these holes in the data don’t let me run my analysis”
- The inferential complications induced by missing data were first addressed in a seminal paper by Rubin (1976, Biometrika)

Biometrika (1976), **63**, 3, *pp.* 581–92 581
Printed in Great Britain

Inference and missing data

BY DONALD B. RUBIN
Educational Testing Service, Princeton, New Jersey

SUMMARY

When making sampling distribution inferences about the parameter of the data, θ, it is appropriate to ignore the process that causes missing data if the missing data are ‘missing at random’ and the observed data are ‘observed at random’, but these inferences are generally conditional on the observed pattern of missing data. When making direct-likelihood or Bayesian inferences about θ, it is appropriate to ignore the process that causes missing data if the missing data are missing at random and the parameter of the missing data process is ‘distinct’ from θ. These conditions are the weakest general conditions under which ignoring the process that causes missing data always leads to correct inferences.

Some key words: Bayesian inference; Incomplete data; Likelihood inference; Missing at random; Missing data; Missing values; Observed at random; Sampling distribution inference.

- Prior to this, some authors had ways of “ignoring” the missing data, but no formal treatment of the *missingness mechanism* existed
- The definitions that Rubin introduced have evolved: see lectures on likelihood-based inference

Missing-Data Mechanisms: Warning

We'll introduce the classification of missing-data mechanisms as they are commonly interpreted, and as presented by Davidian and Tsiatis

However, as we'll see in the lectures on likelihood-based inference, this is not exactly the interpretation that Rubin intended

Missing-Data Mechanisms: Missing Completely at Random

Data are said to be *missing completely at random* (MCAR) if

$$p(R = r \mid z) = p(R = r)$$

Interpreted as

- $R \perp\!\!\!\perp Z$ (R and Z are independent)
- Missingness has nothing to do with the study variables

Missing-Data Mechanisms: Missing Completely at Random

$$\text{MCAR:} \quad p(R = r \mid z) = p(R = r)$$

Example:

let's say $Z = (Sex, Age, Income)$

- Say $r = 110$,

$$p(R = 110 \mid M, 25, 10K) = p(R = 110 \mid F, 70, 60K) = p(R = 110)$$

- Same for all other response patterns r
- We conclude

$$R \perp\!\!\!\perp (Sex, Age, Income)$$

Missing-Data Mechanisms: Missing Completely at Random

$$\text{MCAR:} \quad p(R = r \mid z) = p(R = r)$$

Example:

let's say $Z = (Sex, Age, Income)$

- Say $r = 110$,

$$p(R = 110 \mid M, 25, 10K) = p(R = 110 \mid F, 70, 60K) = p(R = 110)$$

- Same for all other response patterns r
- We conclude

$$R \perp\!\!\!\perp (Sex, Age, Income)$$

Missing-Data Mechanisms: Missing Completely at Random

$$\text{MCAR:} \quad p(R = r \mid z) = p(R = r)$$

Example:

let's say $Z = (Sex, Age, Income)$

- Say $r = 110$,

$$p(R = 110 \mid M, 25, 10K) = p(R = 110 \mid F, 70, 60K) = p(R = 110)$$

- Same for all other response patterns r
- We conclude

$$R \perp\!\!\!\perp (Sex, Age, Income)$$

Missing-Data Mechanisms: Missing Completely at Random

$$\text{MCAR:} \quad p(R = r \mid z) = p(R = r)$$

Example:

let's say $Z = (Sex, Age, Income)$

- Say $r = 110$,

$$p(R = 110 \mid M, 25, 10K) = p(R = 110 \mid F, 70, 60K) = p(R = 110)$$

- Same for all other response patterns r
- We conclude

$$R \perp\!\!\!\perp (Sex, Age, Income)$$

Missing-Data Mechanisms: Missing at Random

Data are said to be *missing at random* (MAR) if

$$p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Interpreted as

- The probability of a response pattern does not depend on the missing data
- The probability of response pattern r as a function of z is constant on $z_{(\bar{r})}$

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example:

let's say $Z = (Sex, Age, Income)$, and only income can be missing

- If $r = 110$,

$$p(R = 110 \mid z) = p(R = 110 \mid z_{(110)}) = p(R = 110 \mid Sex, Age)$$

- If $r = 111$,

$$p(R = 111 \mid z) = p(R = 111 \mid z_{(111)}) = p(R = 111 \mid Sex, Age, Income)$$

- However, since only income can be missing,

$$p(R = 111 \mid z) = 1 - p(R = 110 \mid z)$$

- Therefore $p(R = 111 \mid z) = p(R = 111 \mid Sex, Age)$ and we conclude

$$R \perp\!\!\!\perp Income \mid Sex, Age$$

- (Here we could simply define R as the indicator of missingness for *Income*)

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example:

let's say $Z = (Sex, Age, Income)$, and only income can be missing

- If $r = 110$,

$$p(R = 110 \mid z) = p(R = 110 \mid z_{(110)}) = p(R = 110 \mid Sex, Age)$$

- If $r = 111$,

$$p(R = 111 \mid z) = p(R = 111 \mid z_{(111)}) = p(R = 111 \mid Sex, Age, Income)$$

- However, since only income can be missing,

$$p(R = 111 \mid z) = 1 - p(R = 110 \mid z)$$

- Therefore $p(R = 111 \mid z) = p(R = 111 \mid Sex, Age)$ and we conclude

$$R \perp\!\!\!\perp Income \mid Sex, Age$$

- (Here we could simply define R as the indicator of missingness for *Income*)

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example:

let's say $Z = (Sex, Age, Income)$, and only income can be missing

- If $r = 110$,

$$p(R = 110 \mid z) = p(R = 110 \mid z_{(110)}) = p(R = 110 \mid Sex, Age)$$

- If $r = 111$,

$$p(R = 111 \mid z) = p(R = 111 \mid z_{(111)}) = p(R = 111 \mid Sex, Age, Income)$$

- However, since only income can be missing,

$$p(R = 111 \mid z) = 1 - p(R = 110 \mid z)$$

- Therefore $p(R = 111 \mid z) = p(R = 111 \mid Sex, Age)$ and we conclude

$$R \perp\!\!\!\perp Income \mid Sex, Age$$

- (Here we could simply define R as the indicator of missingness for *Income*)

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example:

let's say $Z = (Sex, Age, Income)$, and only income can be missing

- If $r = 110$,

$$p(R = 110 \mid z) = p(R = 110 \mid z_{(110)}) = p(R = 110 \mid Sex, Age)$$

- If $r = 111$,

$$p(R = 111 \mid z) = p(R = 111 \mid z_{(111)}) = p(R = 111 \mid Sex, Age, Income)$$

- However, since only income can be missing,

$$p(R = 111 \mid z) = 1 - p(R = 110 \mid z)$$

- Therefore $p(R = 111 \mid z) = p(R = 111 \mid Sex, Age)$ and we conclude

$$R \perp\!\!\!\perp Income \mid Sex, Age$$

- (Here we could simply define R as the indicator of missingness for *Income*)

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example:

let's say $Z = (Sex, Age, Income)$, and only income can be missing

- If $r = 110$,

$$p(R = 110 \mid z) = p(R = 110 \mid z_{(110)}) = p(R = 110 \mid Sex, Age)$$

- If $r = 111$,

$$p(R = 111 \mid z) = p(R = 111 \mid z_{(111)}) = p(R = 111 \mid Sex, Age, Income)$$

- However, since only income can be missing,

$$p(R = 111 \mid z) = 1 - p(R = 110 \mid z)$$

- Therefore $p(R = 111 \mid z) = p(R = 111 \mid Sex, Age)$ and we conclude

$$R \perp\!\!\!\perp Income \mid Sex, Age$$

- (Here we could simply define R as the indicator of missingness for *Income*)

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example:

let's say $Z = (Sex, Age, Income)$, and only income can be missing

- If $r = 110$,

$$p(R = 110 \mid z) = p(R = 110 \mid z_{(110)}) = p(R = 110 \mid Sex, Age)$$

- If $r = 111$,

$$p(R = 111 \mid z) = p(R = 111 \mid z_{(111)}) = p(R = 111 \mid Sex, Age, Income)$$

- However, since only income can be missing,

$$p(R = 111 \mid z) = 1 - p(R = 110 \mid z)$$

- Therefore $p(R = 111 \mid z) = p(R = 111 \mid Sex, Age)$ and we conclude

$$R \perp\!\!\!\perp Income \mid Sex, Age$$

- (Here we could simply define R as the indicator of missingness for *Income*)

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example:

let's say $Z = (Sex, Age, Income)$, and any missingness pattern is possible

- If $r = 110$,

$$p(R = 110 \mid z) = p(R = 110 \mid z_{(110)}) = p(R = 110 \mid Sex, Age)$$

- If $r = 111$,

$$p(R = 111 \mid z) = p(R = 111 \mid z_{(111)}) = p(R = 111 \mid Sex, Age, Income)$$

- If $r = 001$,

$$p(R = 001 \mid z) = p(R = 001 \mid z_{(001)}) = p(R = 001 \mid Income)$$

- If $r = 000$,

$$p(R = 000 \mid z) = p(R = 000 \mid z_{(000)}) = p(R = 000)$$

- How do you like this as an assumption?

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example:

let's say $Z = (Sex, Age, Income)$, and any missingness pattern is possible

- If $r = 110$,

$$p(R = 110 \mid z) = p(R = 110 \mid z_{(110)}) = p(R = 110 \mid Sex, Age)$$

- If $r = 111$,

$$p(R = 111 \mid z) = p(R = 111 \mid z_{(111)}) = p(R = 111 \mid Sex, Age, Income)$$

- If $r = 001$,

$$p(R = 001 \mid z) = p(R = 001 \mid z_{(001)}) = p(R = 001 \mid Income)$$

- If $r = 000$,

$$p(R = 000 \mid z) = p(R = 000 \mid z_{(000)}) = p(R = 000)$$

- How do you like this as an assumption?

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example:

let's say $Z = (Sex, Age, Income)$, and any missingness pattern is possible

- If $r = 110$,

$$p(R = 110 \mid z) = p(R = 110 \mid z_{(110)}) = p(R = 110 \mid Sex, Age)$$

- If $r = 111$,

$$p(R = 111 \mid z) = p(R = 111 \mid z_{(111)}) = p(R = 111 \mid Sex, Age, Income)$$

- If $r = 001$,

$$p(R = 001 \mid z) = p(R = 001 \mid z_{(001)}) = p(R = 001 \mid Income)$$

- If $r = 000$,

$$p(R = 000 \mid z) = p(R = 000 \mid z_{(000)}) = p(R = 000)$$

- How do you like this as an assumption?

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example:

let's say $Z = (Sex, Age, Income)$, and any missingness pattern is possible

- If $r = 110$,

$$p(R = 110 \mid z) = p(R = 110 \mid z_{(110)}) = p(R = 110 \mid Sex, Age)$$

- If $r = 111$,

$$p(R = 111 \mid z) = p(R = 111 \mid z_{(111)}) = p(R = 111 \mid Sex, Age, Income)$$

- If $r = 001$,

$$p(R = 001 \mid z) = p(R = 001 \mid z_{(001)}) = p(R = 001 \mid Income)$$

- If $r = 000$,

$$p(R = 000 \mid z) = p(R = 000 \mid z_{(000)}) = p(R = 000)$$

- How do you like this as an assumption?

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example:

let's say $Z = (Sex, Age, Income)$, and any missingness pattern is possible

- If $r = 110$,
$$p(R = 110 \mid z) = p(R = 110 \mid z_{(110)}) = p(R = 110 \mid Sex, Age)$$
- If $r = 111$,
$$p(R = 111 \mid z) = p(R = 111 \mid z_{(111)}) = p(R = 111 \mid Sex, Age, Income)$$
- If $r = 001$,
$$p(R = 001 \mid z) = p(R = 001 \mid z_{(001)}) = p(R = 001 \mid Income)$$
- If $r = 000$,
$$p(R = 000 \mid z) = p(R = 000 \mid z_{(000)}) = p(R = 000)$$
- How do you like this as an assumption?

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example:

let's say $Z = (Sex, Age, Income)$, and any missingness pattern is possible

- If $r = 110$,

$$p(R = 110 \mid z) = p(R = 110 \mid z_{(110)}) = p(R = 110 \mid Sex, Age)$$

- If $r = 111$,

$$p(R = 111 \mid z) = p(R = 111 \mid z_{(111)}) = p(R = 111 \mid Sex, Age, Income)$$

- If $r = 001$,

$$p(R = 001 \mid z) = p(R = 001 \mid z_{(001)}) = p(R = 001 \mid Income)$$

- If $r = 000$,

$$p(R = 000 \mid z) = p(R = 000 \mid z_{(000)}) = p(R = 000)$$

- How do you like this as an assumption?

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example: say $Z = (Z_1, Z_2)$, $\quad (R_1, R_2) \in \{0, 1\}^2$

- $p(R_1 = 0, R_2 = 0 \mid Z_1 = z_1, Z_2 = z_2) = f_{00}$
- $p(R_1 = 1, R_2 = 0 \mid Z_1 = z_1, Z_2 = z_2) = f_{10}(z_1)$
- $p(R_1 = 0, R_2 = 1 \mid Z_1 = z_1, Z_2 = z_2) = f_{01}(z_2)$
- $p(R_1 = 1, R_2 = 1 \mid Z_1 = z_1, Z_2 = z_2) = 1 - f_{00} - f_{10}(z_1) - f_{01}(z_2)$

So MAR in general is NOT a conditional independence statement!

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example: say $Z = (Z_1, Z_2)$, $\quad (R_1, R_2) \in \{0, 1\}^2$

- $p(R_1 = 0, R_2 = 0 \mid Z_1 = z_1, Z_2 = z_2) = f_{00}$
- $p(R_1 = 1, R_2 = 0 \mid Z_1 = z_1, Z_2 = z_2) = f_{10}(z_1)$
- $p(R_1 = 0, R_2 = 1 \mid Z_1 = z_1, Z_2 = z_2) = f_{01}(z_2)$
- $p(R_1 = 1, R_2 = 1 \mid Z_1 = z_1, Z_2 = z_2) = 1 - f_{00} - f_{10}(z_1) - f_{01}(z_2)$

So MAR in general is NOT a conditional independence statement!

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example: say $Z = (Z_1, Z_2), \quad (R_1, R_2) \in \{0, 1\}^2$

- $p(R_1 = 0, R_2 = 0 \mid Z_1 = z_1, Z_2 = z_2) = f_{00}$
- $p(R_1 = 1, R_2 = 0 \mid Z_1 = z_1, Z_2 = z_2) = f_{10}(z_1)$
- $p(R_1 = 0, R_2 = 1 \mid Z_1 = z_1, Z_2 = z_2) = f_{01}(z_2)$
- $p(R_1 = 1, R_2 = 1 \mid Z_1 = z_1, Z_2 = z_2) = 1 - f_{00} - f_{10}(z_1) - f_{01}(z_2)$

So MAR in general is NOT a conditional independence statement!

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example: say $Z = (Z_1, Z_2)$, $\quad (R_1, R_2) \in \{0, 1\}^2$

- $p(R_1 = 0, R_2 = 0 \mid Z_1 = z_1, Z_2 = z_2) = f_{00}$
- $p(R_1 = 1, R_2 = 0 \mid Z_1 = z_1, Z_2 = z_2) = f_{10}(z_1)$
- $p(R_1 = 0, R_2 = 1 \mid Z_1 = z_1, Z_2 = z_2) = f_{01}(z_2)$
- $p(R_1 = 1, R_2 = 1 \mid Z_1 = z_1, Z_2 = z_2) = 1 - f_{00} - f_{10}(z_1) - f_{01}(z_2)$

So MAR in general is NOT a conditional independence statement!

Missing-Data Mechanisms: Missing at Random

$$\text{MAR:} \quad p(R = r \mid z) = p(R = r \mid z_{(r)})$$

Example: say $Z = (Z_1, Z_2)$, $\quad (R_1, R_2) \in \{0, 1\}^2$

- $p(R_1 = 0, R_2 = 0 \mid Z_1 = z_1, Z_2 = z_2) = f_{00}$
- $p(R_1 = 1, R_2 = 0 \mid Z_1 = z_1, Z_2 = z_2) = f_{10}(z_1)$
- $p(R_1 = 0, R_2 = 1 \mid Z_1 = z_1, Z_2 = z_2) = f_{01}(z_2)$
- $p(R_1 = 1, R_2 = 1 \mid Z_1 = z_1, Z_2 = z_2) = 1 - f_{00} - f_{10}(z_1) - f_{01}(z_2)$

So MAR in general is NOT a conditional independence statement!

Missing-Data Mechanisms: Missing Not at Random

Data are said to be *missing not at random* (MNAR) if

$$p(R = r \mid z) \neq p(R = r \mid z_{(r)})$$

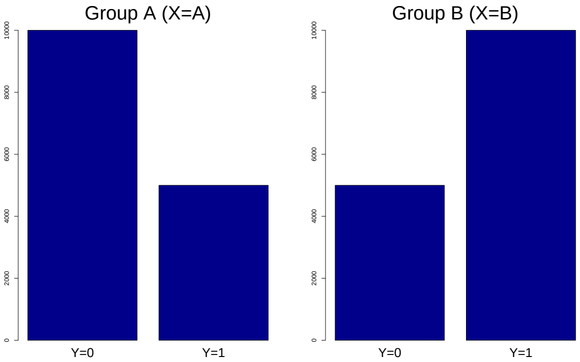
- Quite literally, anything that cannot be written as MAR
- The probability of observing r depends on the components of Z not observed when $R = r$
- Probably the most realistic scenario, and the most difficult to handle

A Toy Example

- $Y \in \{0, 1\}$: indicates presence of a feature, sometimes missing
- $X \in \{A, B\}$: population groups, always observed
- $R \in \{0, 1\}$: response indicator for Y

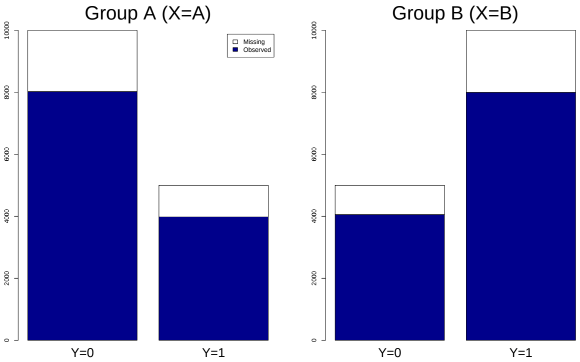
A Toy Example: Full Data

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



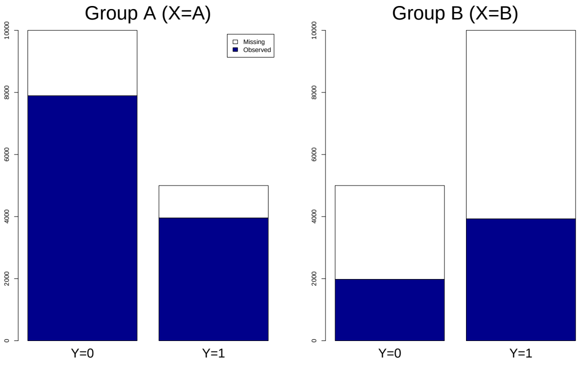
A Toy Example: Missing Completely at Random

$$p(R = 1 \mid x, y) = p(R = 1) = 0.8$$



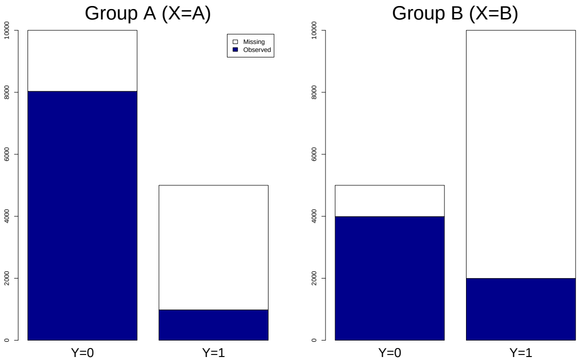
A Toy Example: Missing at Random

$$p(R = 1 \mid x, y) = p(R = 1 \mid x) = 0.8I(x = A) + 0.4I(x = B)$$



A Toy Example: Missing Not at Random

$$p(R = 1 \mid x, y) = p(R = 1 \mid y) = 0.8I(y = 0) + 0.2I(y = 1)$$



What Can We Conclude So Far?

- In general, missing data complicates inference
- In the scale of complication

$$MCAR << MAR <<<<<<<<<<<<<<<< MNAR$$

- But how can we know?
  - MCAR vs MAR?: doable, but relies on assumption that MAR holds
  - MAR vs MNAR?: not possible based on your observed data – MNAR mechanisms depend on data that are not observed
  - The data analyst must adopt an assumption about the mechanism without being able to verify it
- "*if one adopts an assumption of MAR, it must be defensible on scientific, subject matter, and/or practical grounds, because it cannot be validated from the data*" Davidian and Tsiatis
- Inference under MNAR is more realistic but more complicated – we'll look into this towards the end of the course

What Can We Conclude So Far?

- In general, missing data complicates inference
- In the scale of complication

$$MCAR << MAR <<<<<<<<<<<<<<<< MNAR$$

- But how can we know?
  - MCAR vs MAR?: doable, but relies on assumption that MAR holds
  - MAR vs MNAR?: not possible based on your observed data – MNAR mechanisms depend on data that are not observed
  - The data analyst must adopt an assumption about the mechanism without being able to verify it
- "*if one adopts an assumption of MAR, it must be defensible on scientific, subject matter, and/or practical grounds, because it cannot be validated from the data*" Davidian and Tsiatis
- Inference under MNAR is more realistic but more complicated – we'll look into this towards the end of the course

What Can We Conclude So Far?

- In general, missing data complicates inference
- In the scale of complication

$$MCAR << MAR <<<<<<<<<<<<<<<< MNAR$$

- But how can we know?
  - MCAR vs MAR?: doable, but relies on assumption that MAR holds
  - MAR vs MNAR?: not possible based on your observed data – MNAR mechanisms depend on data that are not observed
  - The data analyst must adopt an assumption about the mechanism without being able to verify it
- "*if one adopts an assumption of MAR, it must be defensible on scientific, subject matter, and/or practical grounds, because it cannot be validated from the data*" Davidian and Tsiatis
- Inference under MNAR is more realistic but more complicated – we'll look into this towards the end of the course

What Can We Conclude So Far?

- In general, missing data complicates inference
- In the scale of complication

$$MCAR << MAR <<<<<<<<<<<<<<<< MNAR$$

- But how can we know?
  - MCAR vs MAR?: doable, but relies on assumption that MAR holds
  - MAR vs MNAR?: not possible based on your observed data – MNAR mechanisms depend on data that are not observed
  - The data analyst must adopt an assumption about the mechanism without being able to verify it
- "*if one adopts an assumption of MAR, it must be defensible on scientific, subject matter, and/or practical grounds, because it cannot be validated from the data*" Davidian and Tsiatis
- Inference under MNAR is more realistic but more complicated – we'll look into this towards the end of the course

Never Work Under MAR?

Most approaches for inference with missing data assume MAR

- Option 1: "don't worry about how the sausage gets made, just eat the sausage!," or the approach of the horse with blinders:

Never Work Under MAR?

Most approaches for inference with missing data assume MAR

- Option 2: you can argue that MAR is not unreasonable. For example, do you have sufficiently rich information that is always observed?
  - Say $Z = (Z_1, Z_2)$
  - Z_1: a vector subject to missingness
  - Z_2: fully observed
  - R: response indicator for Z_1
  - MAR: $\quad p(R = r \mid z_1, z_2) = p(R = r \mid z_{1(r)}, z_2)$
  - If assuming $p(R = r \mid z_1, z_2) = p(R = r \mid z_2)$ is reasonable, then MAR is reasonable because MAR is more general

Never Work Under MAR?

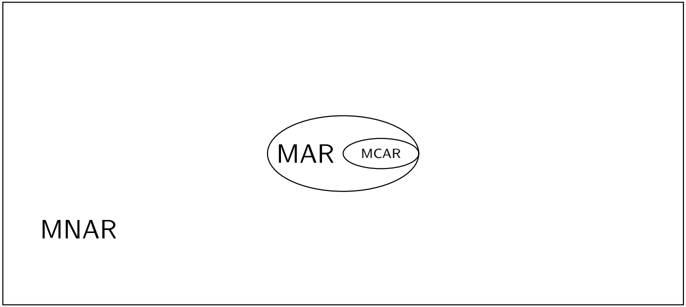
Most approaches for inference with missing data assume MAR

- Option 3: take this class, think about these issues, contribute to creating better solutions!

Summary

Main take-aways from today's lecture:

- Proper handling of missing data requires proper notation
- Universe of missing-data assumptions:



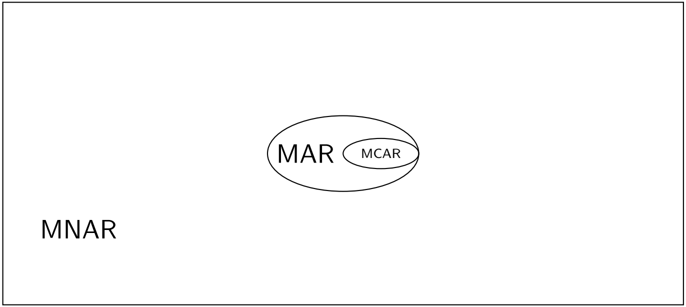
Next lecture:

- Naïve methods for handling missing data: imputation and complete cases
- Reading: Chapter 2 in Davidian and Tsiatis

Summary

Main take-aways from today's lecture:

- Proper handling of missing data requires proper notation
- Universe of missing-data assumptions:



Next lecture:

- Naïve methods for handling missing data: imputation and complete cases
- Reading: Chapter 2 in Davidian and Tsiatis