Missing Data: What are the Options?

Shaun Seaman

MRC Biostatistics Unit, Cambridge

Overview

- Missing data: how arise and why a problem
- Missing data mechanisms
- When is complete-case analysis valid?
- Inverse probability weighting
- Multiple imputation
- Full-likelihood methods
- Doubly-robust methods
- Two special interests

Missing Data and How They Arise

Missing data common in epidemiological research

- Subjects miss scheduled visits
- Drop out of study
- Decline to answer questions
- Refuse to participate
- Biological samples lost or tests fail

Digression: a broader concept is *coarsened* data

X is coarsened if we observe $\mathcal{A}$ such that $X \in \mathcal{A}$

- censored data: $X \in [c, \infty)$
- discretised data: $X \in [a, b]$

Missing data complicate the analysis

Good advice is not to have any! But not practical

Why Missing Data are a Problem

Example 1: Bias

Estimate average weight in population

Take sample and measure weights

Scales only go up to 120Kg

Nurse enters 'NA' if maximum weight exceeded

Example 2: Inefficiency

Estimate parameters of model for relation between blood glucose and smoking, BMI, social class,...

Blood glucose observed for everyone

98% of covariate values observed

But 50% of subjects have at least one covariate missing

Missing Data Mechanisms (MDMs)

MDM is assumption about what missingness can depend on

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

Let $\boldsymbol{X}$ be a vector of variables measured on each individual

Let $\boldsymbol{R}$ denote missingness pattern of $\boldsymbol{X}$

$$\boldsymbol{X} = (1, 3, 5, \text{NA}) \qquad \boldsymbol{R} = (1, 1, 1, 0)$$
$$\boldsymbol{X} = (\text{NA}, \text{NA}, 4.2, 1.9) \qquad \boldsymbol{R} = (0, 0, 1, 1)$$

Write $\boldsymbol{X} = (\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}})$

MCAR : $P(\boldsymbol{R} = \boldsymbol{r} \mid \boldsymbol{X}) = P(\boldsymbol{R} = \boldsymbol{r})$
MAR : $P(\boldsymbol{R} = \boldsymbol{r} \mid \boldsymbol{X}) = P(\boldsymbol{R} = \boldsymbol{r} \mid \boldsymbol{X}_{\text{obs}})$
MNAR : not MAR

Monotone missing data

Suppose $\boldsymbol{X} = (X_1, X_2, X_3)$ and X_j is observed only if X_{j-1} is observed

Only possible values of $\boldsymbol{R}$ are

$$\boldsymbol{R} = (1,1,1) \qquad \boldsymbol{R} = (1,1,0)$$
$$\boldsymbol{R} = (1,0,0) \qquad \boldsymbol{R} = (0,0,0)$$

Monotone missingness in longitudinal studies if missingness all due to dropout

$$\begin{aligned} P[\boldsymbol{R} = (1,1,1) \mid \boldsymbol{X}] &= P(R_1 = 1 \mid \boldsymbol{X}) \times P(R_2 = 1 \mid R_1 = 1, \boldsymbol{X}) \\ &\quad \times P(R_3 = 1 \mid R_2 = 1, \boldsymbol{X}) \\ P[\boldsymbol{R} = (1,1,0) \mid \boldsymbol{X}] &= P(R_1 = 1 \mid \boldsymbol{X}) \times P(R_2 = 1 \mid R_1 = 1, \boldsymbol{X}) \\ &\quad \times P(R_3 = 0 \mid R_2 = 1, \boldsymbol{X}) \end{aligned}$$

MCAR: $P(R_j = 1 \mid R_{j-1} = 1, \boldsymbol{X}) = P(R_j = 1 \mid R_{j-1} = 1)$

MAR : $P(R_j = 1 \mid R_{j-1} = 1, \boldsymbol{X}) = P(R_j = 1 \mid R_{j-1} = 1, X_1, \ldots, X_{j-1})$

Missing data mechanism depends on the set of variables

What does "$\boldsymbol{X}$ is MAR" mean?

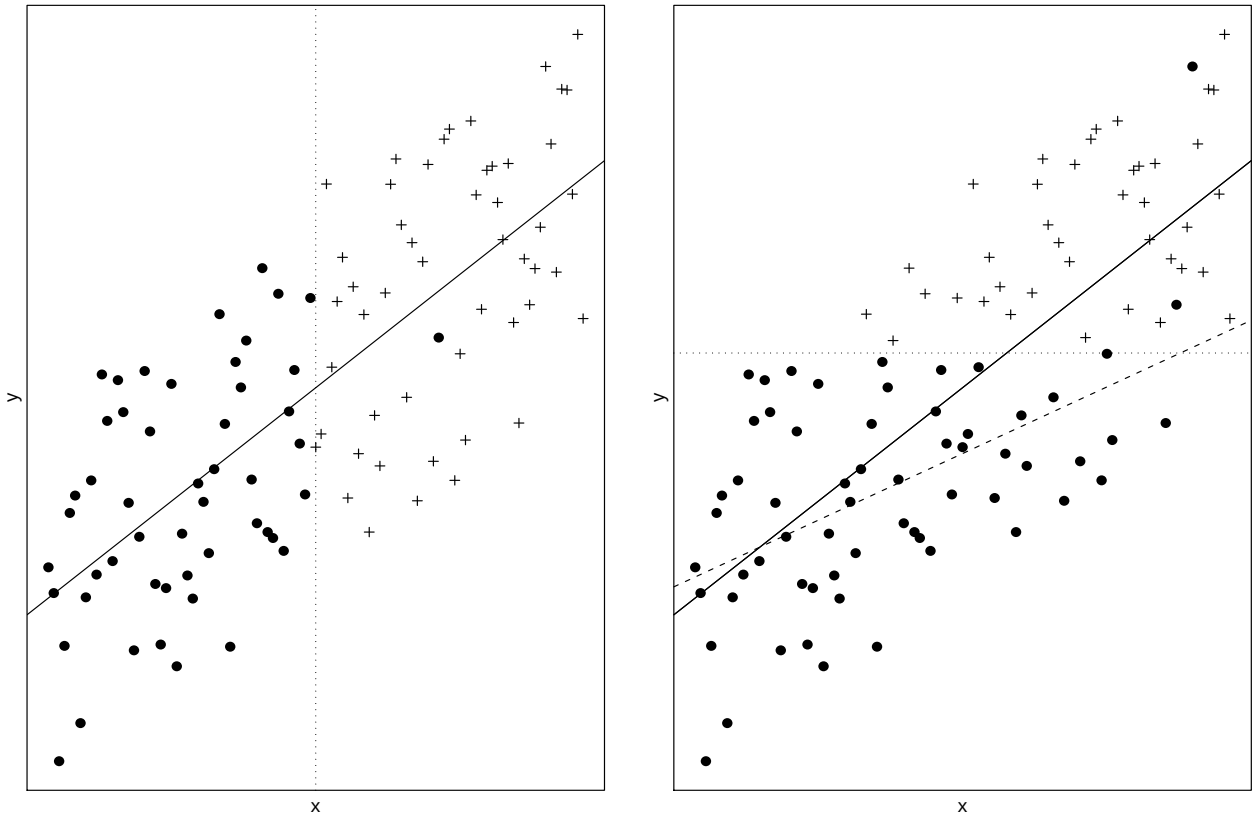
- If $\boldsymbol{X}$ = blood pressure
  MAR :
  whether blood pressure is observed is independent of blood pressure
- If $\boldsymbol{X}$ = blood pressure and fully-observed variables social class, sex, age, ...
  MAR:
  whether blood pressure is observed is independent of blood pressure given social class, sex, age, ...

Including additional variables can make MAR assumption more plausible

When is Complete-Case Analysis Valid?

CC analysis is valid if complete cases are representative of whole sample

More specifically, same relation between variables described by analysis model



Inverse Probability Weighting (IPW)

let $\boldsymbol{H}$ be set of variables that predict whether a complete case

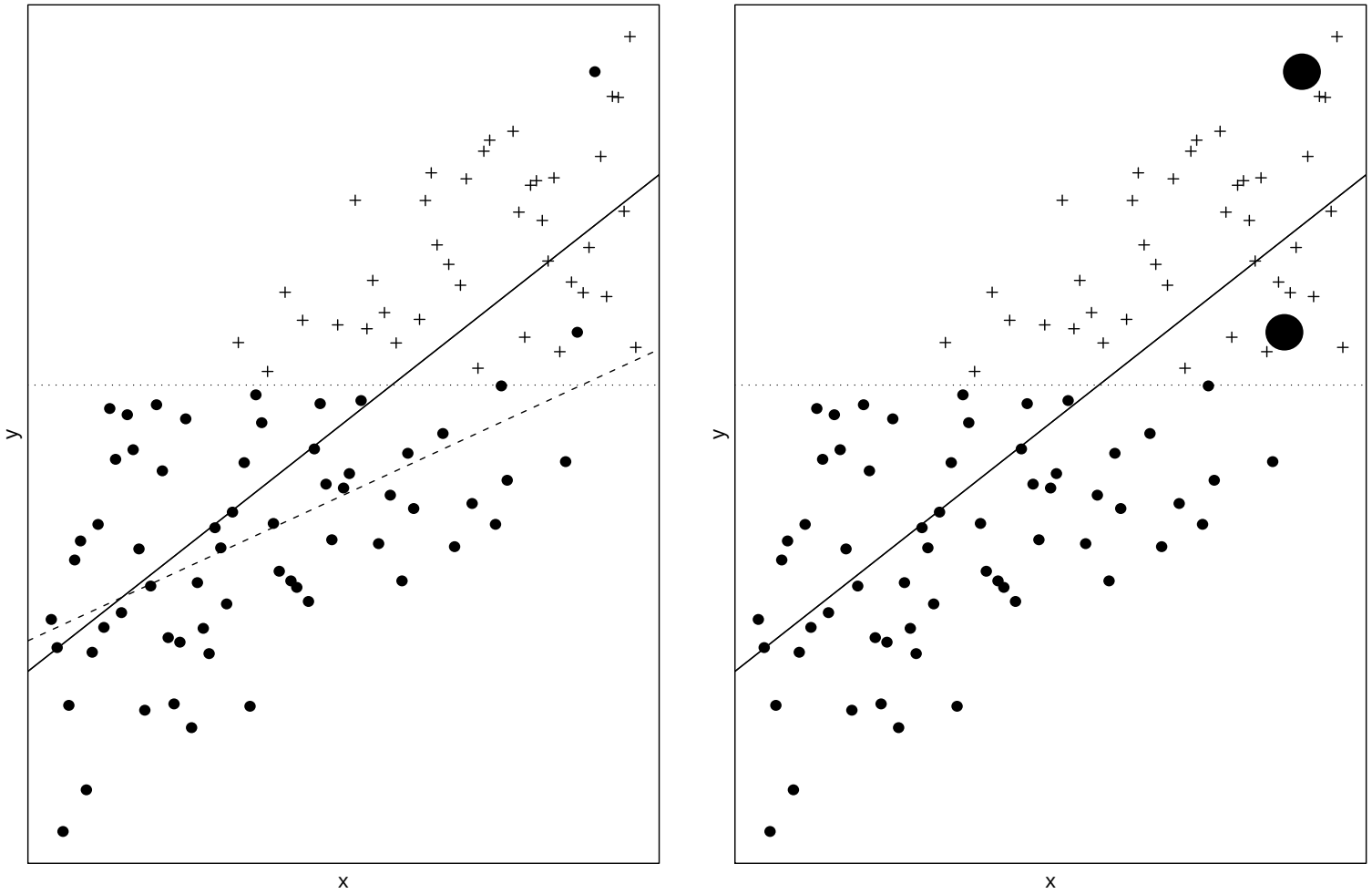
Let $\pi(\boldsymbol{H}) = P(\text{CC} \mid \boldsymbol{H})$

Weight each complete case by $1/\pi(\boldsymbol{H})$

IPW gives valid inference if $P(\text{CC} \mid \boldsymbol{H}, \boldsymbol{X}) = P(\text{CC} \mid \boldsymbol{H})$, where $\boldsymbol{X}$ are variables in the analysis model

$\pi(\boldsymbol{H})$ usually unknown, so model specified and estimated using observed data

Also used for estimating treatment effects in non-randomised studies


y
x
y
x

Multiple Imputation (MI)

	X_1	X_2	X_3
1	1	10	20.1
2	1	8	15.2
3	0	?	4.7
4	0	2	3.9
5	1	?	?
6	0	-1	?

1	10	20.1		1	10	20.1
1	8	15.2		1	8	15.2
0	2.3	4.7		0	1.9	4.7
0	2	3.9		0	2	3.9
1	9.3	16.9		1	10.5	19.3
0	-1	-2.3		0	-1	-1.7

- Specify model for joint distribution of $\boldsymbol{X}$
- Fit model to observed data
- For each individual, sample $\boldsymbol{X}_{\text{mis}}$ using fitted model and $\boldsymbol{X}_{\text{obs}}$
- Generate multiple complete datasets
- Analyse each complete dataset
- Combine estimates (usually Rubin's Rules)

MI is valid under MAR, because

MAR implies $p(\boldsymbol{X}_{\text{mis}} \mid \boldsymbol{X}_{\text{obs}}, \boldsymbol{R}) = p(\boldsymbol{X}_{\text{mis}} \mid \boldsymbol{X}_{\text{obs}})$

So, can

- impute $\boldsymbol{X}_{\text{mis}}$ using $p(\boldsymbol{X}_{\text{mis}} \mid \boldsymbol{X}_{\text{obs}})$, and
- estimate $p(\boldsymbol{X}_{\text{mis}} \mid \boldsymbol{X}_{\text{obs}})$ from the observed data

IPW versus MI

- IPW assumes model for probability of being a complete case
- MI assumes model for joint distribution of data
- MI more efficient if some incomplete cases have some data
- IPW may be preferred if incomplete cases provide little information

Full-Likelihood Methods

Similar to MI

Specify model for joint distribution of $\boldsymbol{X}$, of which analysis model is part

E.g. analysis model is linear regression

$$Y_i = \beta_0 + \beta_1 Z_i + \epsilon_i$$

If specify distribution for Z, then have joint distribution for (Z, Y)

Fit this model to (Z, Y)

Extract estimates of parameters of analysis model

Like MI, valid if data are MAR

MI often easier to do in practice, especially if want to include auxiliary variables

Doubly Robust Methods (DR)

IPW requires model for probability of being complete case

MI requires model for joint distribution of data

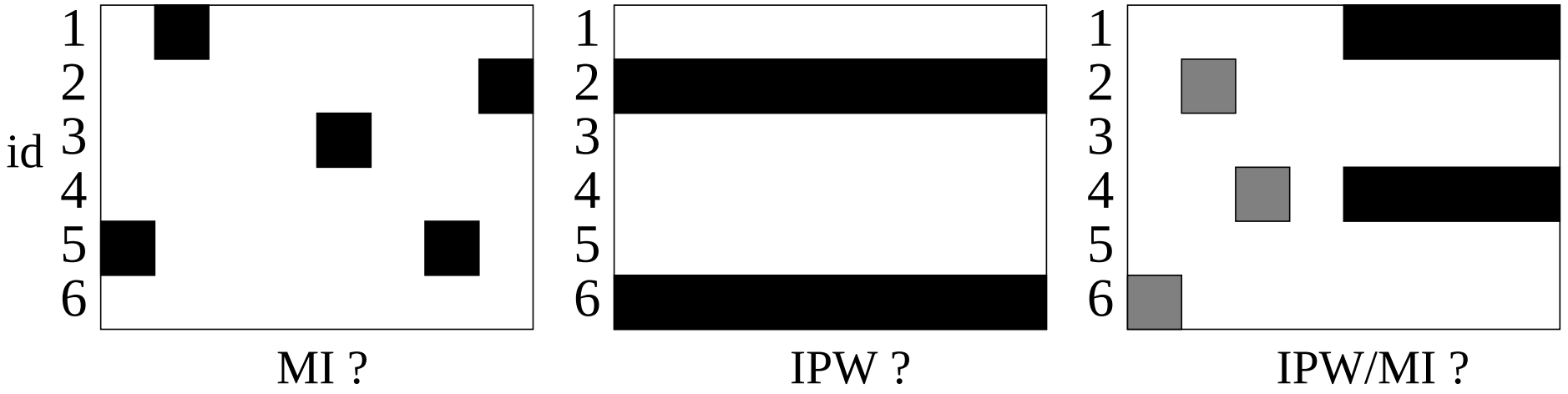
DR uses both models and is valid if either is correct

If both models correct, DR is

- more efficient than IPW, but
- less efficient than MI

Limited utility so far

Combining IPW and MI



IPW/MI also for surveys with unequal sampling fractions and missing data

IPW/MI has been used (e.g. Priebe et al., 2004; Stansfeld et al., 2008)

Application to 1958 British Birth Cohort

- 17638 individuals born in UK during one week in 1958
  16334 still alive at age 45
  9014 (55%) participated in biomedical survey
- Thomas et al. (2007) regressed blood glucose (high/normal) at age 45 on BMI and waist size at age 45 and variables measured at birth
  7518 had observed glucose, BMI and waist size
  Thomas et al. imputed missing birth variables using MICE
- But are 7518 representative of 16334?
  We used IPW/MI
- Need model for P(glucose, BMI & waist size complete)
  Used predictors measured at birth, age 7 and 11 (e.g. birth weight, class)
  Disadvantaged less likely to be complete for glucose, BMI & waist size

Missing Predictors of Missingness in IPW

IPW requires

- Set of variables $\boldsymbol{H}$ that predict whether a complete case
- Model for $\pi(\boldsymbol{H}) = P(\text{CC} \mid \boldsymbol{H})$

Easy to fit this model and estimate $\pi(\boldsymbol{H})$ if $\boldsymbol{H}$ is fully observed

But what if $\boldsymbol{H}$ can be partially missing?

- Crude methods, e.g. replace missing $\boldsymbol{H}$ by mean
- Only use fully-observed $\boldsymbol{H}$
- Assume $P(\text{CC})$ depends only on observed $\boldsymbol{H}$ (D'Agostino & Rubin, 2000)
- Multiply impute missing $\boldsymbol{H}$ (Qu & Lipkovich, 2009)

Summary

- Missing data
- IPW, MI, full-likelihood and DR
- Combining IPW and MI
- IPW with missing predictors of missingness

Predictor of	Thomas		IPW/MI		All MI	
High Glucose	log OR	SE	log OR	SE	log OR	SE
short gestation	0.45	0.22	0.46	0.23	0.44	0.20
pre-eclampsia	**0.47**	0.26	**0.55**	0.27	**0.47**	0.25
smoke in pregncy	0.02	0.14	0.04	0.14	0.04	0.14
mother overwt	**0.28**	0.15	**0.35**	0.15	**0.18**	0.12
manual at birth	0.37	0.17	0.44	0.18	0.39	0.17
low birth weight	−0.30	0.09	−0.30	0.09	−0.32	0.09
BMI age 45	0.04	0.02	0.02	0.02	0.03	0.02
waist (cm) age 45	0.07	0.01	0.07	0.01	0.07	0.01

Stronger relation between glucose and pre-eclampsia/mother overwt in disadvantaged than in advantaged

When use All MI, relation similar in individuals with imputed glucose

Ordinary IPW uses only 5673 (75% of 7518) with complete birth data