

Multiple imputation for missing data in life course studies

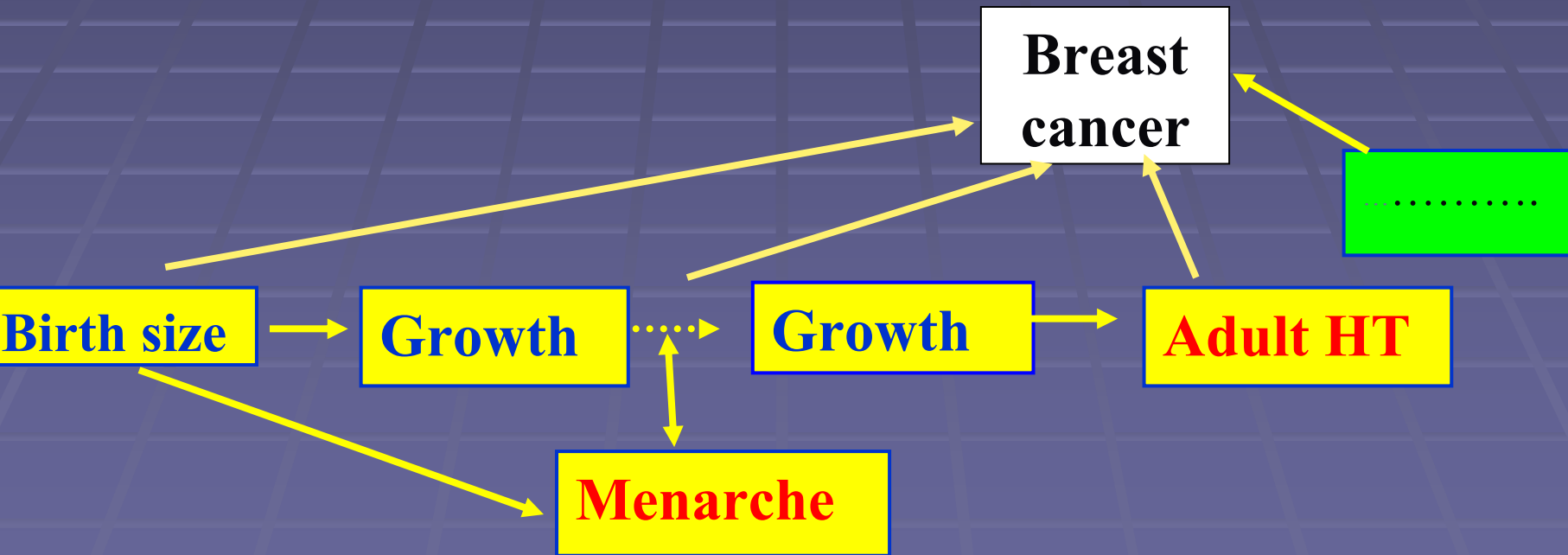
Bianca De Stavola and Valerie McCormack
(London School of Hygiene and Tropical Medicine)

- Motivating example
- Types of missingness and common strategies
- Multiple imputation
- A suite of MI programs

Motivating example

Breast cancer aetiology:

- Several established risk factors (adult HT, menarche)
- New focus on early life and childhood growth



MRC 1946 birth cohort (N=2187):

repeated anthropometric measures in childhood

<i>Childhood height measured at:</i>	N	% missing
2 yrs	1782	18.5
4 yrs	1944	11.1
7 yrs	1925	12.0
11 yrs	1862	14.9
15 yrs	1689	22.8
ALL	904	41.3

Pattern and type of missingness

Data: $Y = (y, X_1, \dots, X_p) = (Y_{\text{obs}}, Y_{\text{mis}})$

	y	X_1	X_2	...	X_p
1			•		
2					
3		•		•	
4			•		
...			•	•	•
...					•
n	•				

MCAR:

$$\Pr(\text{missing}) = \text{not } f(Y_{\text{obs}}, Y_{\text{mis}})$$

MAR:

$$\Pr(\text{missing}) = f(Y_{\text{obs}})$$

NMAR:

$$\Pr(\text{missing}) = f(Y_{\text{obs}}, Y_{\text{mis}})$$

Strategies

- 1. Analyse only those with complete data**
- 2. Available case analysis**
- 3. Inclusion of a “missing value” category**
- 4. Use methods not requiring complete data**
- 5. Replacing missing value with imputed**

Strategies

- ~~1.~~ Analyse only those with complete data
- ~~2.~~ Available case analysis
- 3. Inclusion of a “missing value” category

Biased even when data are MCAR

(Greenland and Finkle 1995)

confounder	RR_E
Level 1	1.45
Level 2	2.03
NK	1.51
overall	1.75

Strategies

- ~~1.~~ Analyse only those with complete data
- ~~2.~~ Available case analysis
- ~~3.~~ Inclusion of a “missing value” category
- 4. Use methods not requiring complete data
- 5. Replacing missing value with imputed

5 - Imputations

If MAR:

Idea: replace missing values with a “guess”

Analysis: same as with complete data

Two types, many variants:

I. SINGLE IMPUTATION

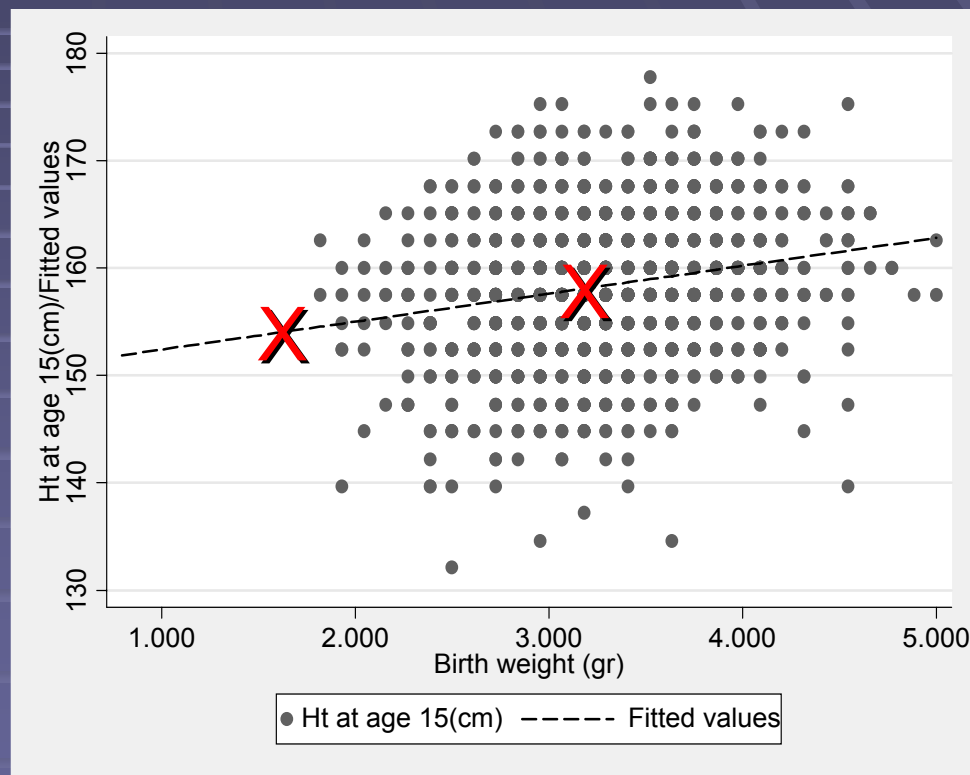
II. MULTIPLE IMPUTATION

I - SINGLE IMPUTATION

a) from a regression model:

replace missing values
with predicted

not good:
↓ true data variation

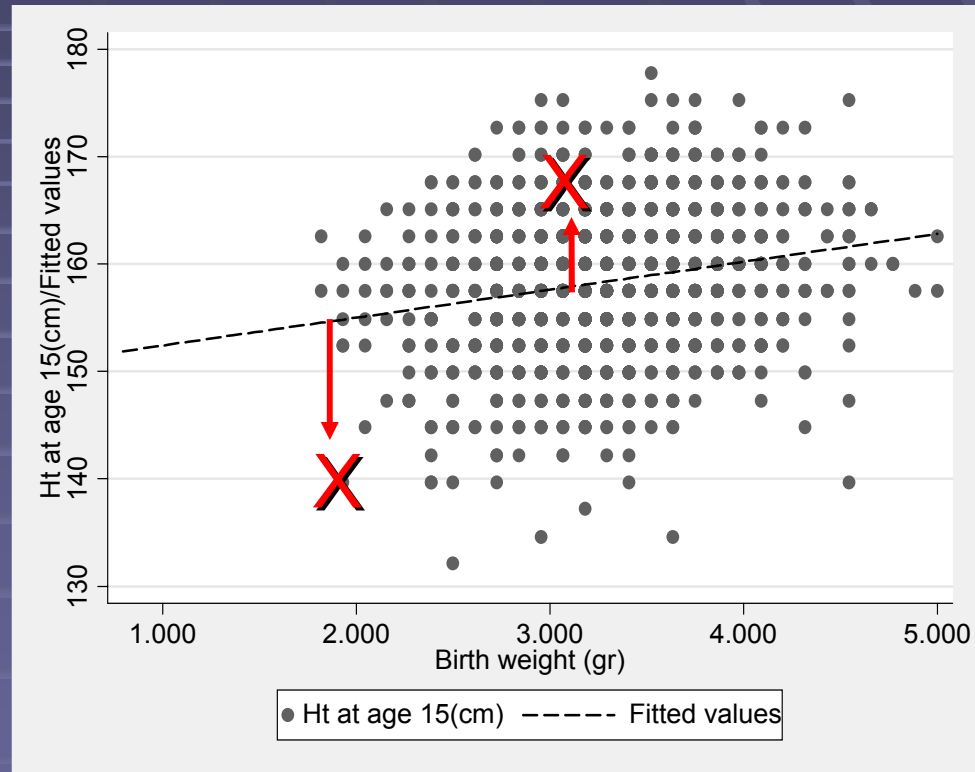


impute.ado

SINGLE IMPUTATION

b) predicted value + random term

Size of random term depends on residual variance of the model



UNSATISFACTORY:

Still pretending the data are observed!

SINGLE IMPUTATION

c) Hot-deck

Replaces record with any missing values with another, but complete, selected at random

Not recommended if several records are incomplete

II - MULTIPLE IMPUTATION

Not one but several data sets are created

- Each has a different set of random draws to replace missing value
- Separate analyses on each data set
- Results summarised

PROBLEM:

generating a 'proper' predictive distribution



More technically.....

- MI replacements are simulated draws from a predictive distribution of the missing data:

$$Y_{\text{mis}}^* \sim P(Y_{\text{mis}} | Y_{\text{obs}}, \theta^*)$$

$$\text{where } \theta^* \sim P(\theta | Y_{\text{obs}})$$

- Require a model for the complete data

$$P(Y | \theta)$$

- **Proper**, i.e. reflect uncertainty about missing data and the parameters

(Shafer, Multiple imputation: a primer. *Stat Methods in Medical Research*, 1999)

MI: three steps

A. Imputation of plausible values:

- Missing values replaced by imputed
- **m** times

B. Analysis of the imputed datasets

C. Combination of the results

B. Analysis

- Each dataset is analysed in the same way:
e.g. : logistic regression
- Save :
 - Point estimates of the statistics of interest:
 $\log(\text{OR}) = Q_{(l)}$
 - Their variance matrix: $U_{(l)}$
- All stored for $l=1,2,\dots,m$

C. Combination

Take the m sample estimates Q_j and variance U_j

For one parameter:

- Overall estimator: Mean (Q_j)
- Its variance: Mean(U_j) + $(1+1/m)$ Var(Q_j)

For k parameters:

- Overall estimators: Mean (Q_j)
- Variance matrix : $(1+ r_1)$ Mean (U_j)

$$r_1 = (1+1/m) \text{trace}[\text{var}(Q_j) (\text{mean}(U_j)^{-1})] / k$$

A. Imputation

Most difficult part

Say x_1 has missing values. Approaches:

- i. Use draws from available observations of x_1
(unconditional draws)
- ii. Use draws from regression models of x_1
(conditional draws)
- iii. [Hot-deck imputation]
- iv. [Markov Chain Monte Carlo techniques]

i) unconditional draws

$$x_{1i} \sim N(\mu, \sigma^2), i=1, \dots, N$$

only $x_{11} x_{12} \dots x_{1a}$ observed, for $a < N$: (\bar{x} obs)



For imputation run !

1. Draw $\sigma^2_{(i)}$ f

2. D

3

4.

Not good for MAR

should condition on:

- Factors affecting x_1
- factors influencing missingness

$$(\mu_{(i)}, \sigma^2_{(i)})$$

St observed $x_{11} x_{12} \dots x_{1a}$ plus imputed in

ii) Simple conditional draws

Assume $x_{1i} \sim N(\beta_0 + \beta_1 x_{2i}, \sigma^2)$,

x_{2i} always observed, $x_2 \rightarrow$ missing mechanism, $X = [1 \ x_2]$,

For imputation run $l = 1, \dots, m$:

1. Draw $\sigma^2_{(l)}$ from $(a-2) S^2_{\text{obs}} / \chi^2_{(a-2)}$
2. Draw $(\beta_{0(l)}, \beta_{1(l)})$ from $N((\hat{\beta}_0, \hat{\beta}_1), \sigma^2_{(l)} (X'X)^{-1})$
3. Draw missing values from $N(\beta_{0(l)} + \beta_{1(l)} x_i, \sigma^2_{(l)})$
4. New dataset: observed plus imputed in step 3

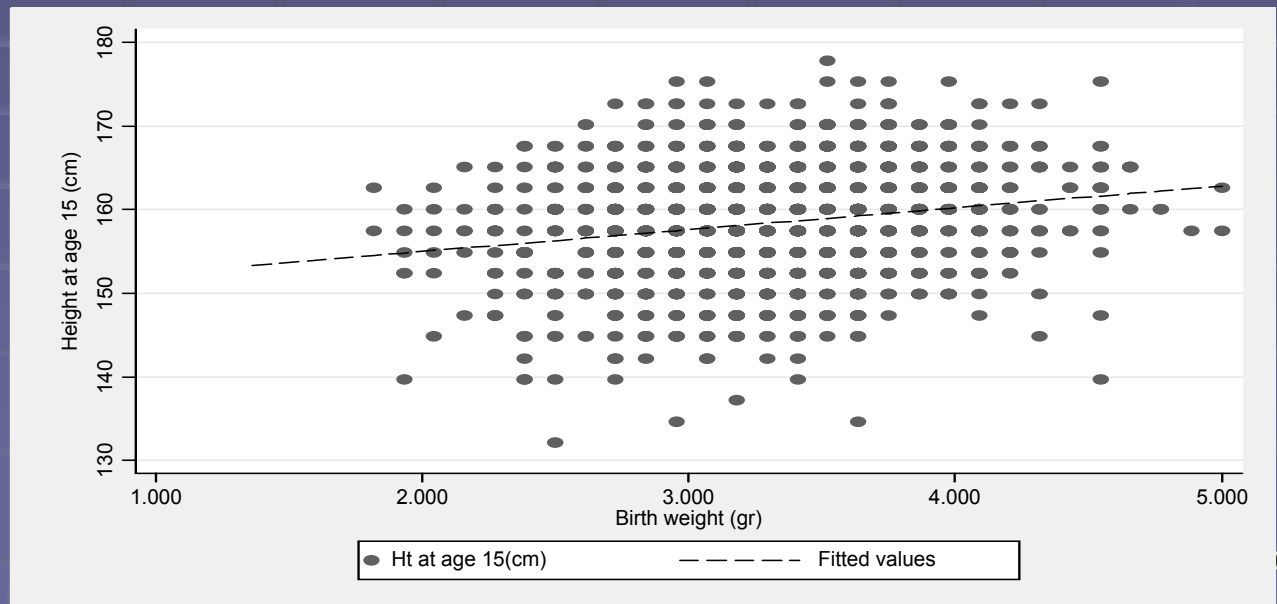
Example

MRC 1946 birth cohort: **2187** women

x_1 : HT at 15 and breast cancer by age 53: **1689**

MI procedure:

- **HT15** = $f(\text{birth weight})$
- **Prob(missing)** = $f(\text{birth weight}, \text{breast cancer})$



MI programs:

A. mi_create_reg.ado

draws missing HT15 using results from regression of observed HT15 on (BW, BRCA) m times to create m imputed datasets

B. mi_logit.ado

runs logistic regression on each imputed dataset and saves the results

C. mi_summary.ado

summarises the results as in Shafer (1997)

Draws for ht at age 15 cond. on BW and BRCA:

Original estimates:

$$\hat{\sigma} = 6.19, \hat{\beta}_0 = 149.79,$$

$$\hat{\beta}_1 = 2.60, \hat{\beta}_2 = 1.48$$

(N=1683)

(l)	$\sigma_{(l)}$	$\beta_{0(l)}$	$\beta_{1(l)}$	$\beta_{2(l)}$
1	6.20	149.75	2.60	1.54
2	6.34	149.91	2.60	1.40
3	6.26	149.43	2.59	1.62
4	6.20	149.74	2.60	1.41
5	6.03	149.83	2.61	1.52

OBSERVED

MI

OR (95% CI)

OR

95% CI.

1.045 (1.00, 1.10)

1.044

(1.00, 1.09)

iii) conditional draws from a random effects model

- Using all childhood growth data \underline{y}_i ($p \times 1$ vector):

p Observed values: $\underline{y}_i = \mathbf{Z} \underline{\eta}_i + \underline{\varepsilon}_i$

q Latent factors: $\underline{\eta}_i = \beta \underline{\mathbf{X}}_i + \underline{\mathbf{u}}_i$

$\underline{\varepsilon} \sim \mathbf{N}(0, \Sigma)$, $\underline{\mathbf{u}} \sim \mathbf{N}(0, \Psi)$, independence assumptions

Explanatory variables: $\underline{\mathbf{X}}_i$

Loading factors (fcn of observation times): \mathbf{Z}

i.e. $y_i \sim \mathbf{N}(\beta \underline{\mathbf{X}}_i, \mathbf{Z}_i \Psi \mathbf{Z}_i' + \Sigma)$

Imputation procedure in similar steps

For imputation run $l=1, \dots, m$:

1. Draw $\Sigma_{(l)}$ from inverse Wishart based on $\hat{\Sigma}$
2. Draw $\Psi_{(l)}$ from inverse Wishart based on $\hat{\Psi}$
3. Draw $\underline{\eta}_{(l)}$ from $N(\underline{\eta}_{\text{pred}}, \mathbf{Z}'\Psi_{(l)}\mathbf{Z})$
4. Draw missing values from $N(\underline{\eta}_{(l)}, \Sigma_{(l)})$
5. New dataset: observed plus imputed in step 3

MI program:

A. mi_create_growth.ado

Logistic regression with imputed growth variables

Use conditional draws, with several explanatory factors (including breast cancer)

		Observed data (N=904, D=33)		Observed and imputed data (N=2187, D=59)	
HEIGHT	Units	OR	95%CI	OR	95%CI
Intercept at 2yrs	<i>cm</i>	1.08	0.71,1.66	1.18	0.87,1.60
Velocity 2-4 yrs	<i>cm/yr</i>	1.02	0.67,1.56	1.14	0.88,1.51
Velocity 4-7 yrs	<i>cm/yr</i>	1.53	1.04,2.24	1.41	1.08,1.85
Velocity 7-11yrs	<i>cm/yr</i>	1.44	0.92,2.25	1.15	0.81,1.62
Velocity 11-15ys	<i>cm/yr</i>	1.23	0.78,1.93	1.30	0.99,1.70
Velocity 15-adulthood	<i>cm/yr</i>	1.05	0.70,1.58	0.94	0.71,1.24

Summary

- **MI requires great care in creating imputed values**
 - A. `mi_create_reg.ado` & `mi_create_growth.ado`
 - B. `mi_logit.ado` & `mi_ologit.ado`
 - C. `mi_summary.ado`
- **Other Stata programs:**
 - `impute.ado`
 - `regmsng.ado`
 - `hotdeck.ado`

 - `implogit.ado`
 - Gary Kings' programs: clarify
 - Ken Scheve's programs