JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Lecture 14

Missing data considerations

Objectives

- Throughout the course we have been sub-setting our data such that we are only including rows of data with non-missing outcomes and exposures.

- Today we will start to explore the possible implications of this practice and think about the underlying assumptions we are making when we do this.

- Upon completion of this session, you will be able to do the following:
    - Define mechanisms that generate missing data *
    - Describe the impact of conducting analyses on complete cases or available data under the different missing data mechanisms *
    - Describe imputation procedures to account for missing data *

Missing data

- Can occur in cross-sectional studies and longitudinal studies

- Why do we see missing data?
  - Persons elect to not respond to all questions
  - Persons are physically/mentally unable to complete assessments
  - Persons no longer wish to participate in a study → drop out
  - Errors in processing biological samples
  - Errors in data entry

- Missing data can occur in for the outcome of interest or for exposures/confounders/effect modifiers/etc.

Missing data mechanisms

- In statistics, we think of the ways in which the missing data was generated
  - We think of the model or probability distribution that generated the missing data.
  - There are 3 types of missing data mechanisms.

Think of a longitudinal data setting, where you have data Y for an individual.

- $\underbrace{Y_i = \begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{ini} \end{bmatrix}} = \begin{bmatrix} Y^o \\ Y^m \end{bmatrix}$ $\begin{matrix} \to \text{observe} \\ \to \text{missing} \end{matrix}$

- $R_i = \begin{bmatrix} R_{i1} \\ \vdots \\ R_{ini} \end{bmatrix}$ is a vector of indicators for whether observation *j* from subject *i* is missing (1) or observed (0)

Y_i
↓

- Consider the joint distribution $f(Y_i, R_i | \theta, \varphi) = \underbrace{\Pr(R_i | Y_i, \varphi) f(Y_i | \theta)}$

↑
R_i

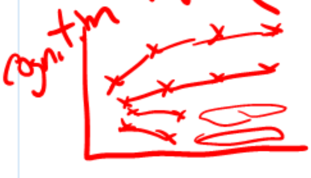
Missing completely at random

$$\Pr(R_i \mid Y_i, \phi) = \Pr(R_i \mid \underbrace{\phi})$$

↳ missing data pattern does not depend on any values of Y, either observed Y^o or missing Y^m

Missing at random

$$\Pr(R_i \mid Y_i, \phi) = \Pr(R_i \mid Y_i^o, Y_i^m, \phi)$$



$$\longrightarrow = \Pr(R_i \mid \underset{\longrightarrow}{Y_i^o}, \phi)$$

=) selection procedure

=) observed data is a specific selection of all possible data

=) missing completely at random within subsets of individuals with the same Y_i^o

Missing not at random

$$\Pr(\underbrace{R_i} \mid Y_i, \phi) = \Pr(\underline{\underline{R_i}} \mid \underset{\sim}{Y_i^o}, \underline{Y_i^m}, \phi)$$

→ Drop out/missing data depends on what I would observe at that assessment had we had access to the participant

Ignorable vs. non-ignorable missing data

The likelihood of the observed data, i.e. (Y^o, R), is:

$$L(\theta, \phi | Y^o, R) = \int_{Y^m} f(Y^o, Y^m, R | \theta, \phi) dY^m = \int_{Y^m} f(Y^o, Y^m | \theta) Pr(R | Y, \phi) dY^m$$

score equation for the full data likelihood

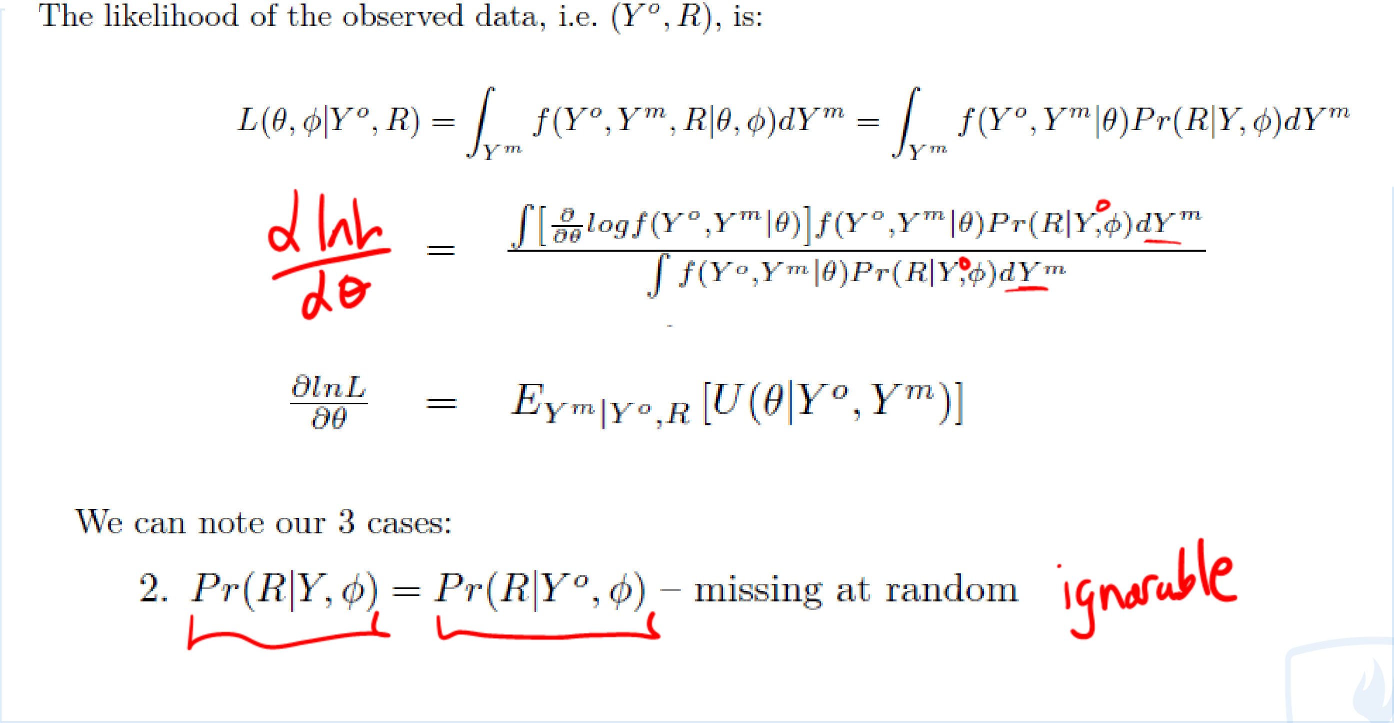
$$\frac{d \ln L}{d\theta} = \frac{\int [\frac{\partial}{\partial\theta} log f(Y^o, Y^m | \theta)] f(Y^o, Y^m | \theta) \boxed{\cancel{Pr(R|Y,\phi)}} dY^m}{\int f(Y^o, Y^m | \theta) \boxed{\cancel{Pr(R|Y,\phi)}} dY^m}$$

$$\frac{\partial ln L}{\partial\theta} = E_{Y^m | Y^o, R}[U(\theta | Y^o, Y^m)]$$

We can note our 3 cases:

1. $Pr(R|Y,\phi) = Pr(R|\phi)$ – missing completely at random ignorable

Ignorable vs. non-ignorable missing data



The likelihood of the observed data, i.e. (Y^o, R), is:

$$L(\theta,\phi|Y^o,R) = \int_{Y^m} f(Y^o,Y^m,R|\theta,\phi)dY^m = \int_{Y^m} f(Y^o,Y^m|\theta)Pr(R|Y,\phi)dY^m$$

$$\frac{d\ln L}{d\theta} = \frac{\int\left[\frac{\partial}{\partial\theta}log f(Y^o,Y^m|\theta)\right]f(Y^o,Y^m|\theta)Pr(R|Y^{o},\phi)d\underline{Y}^m}{\int f(Y^o,Y^m|\theta)Pr(R|Y^{o},\phi)d\underline{Y}^m}$$

$$\frac{\partial lnL}{\partial\theta} = E_{Y^m|Y^o,R}\left[U(\theta|Y^o,Y^m)\right]$$

We can note our 3 cases:

2. $\underbrace{Pr(R|Y,\phi)} = \underbrace{Pr(R|Y^o,\phi)}$ – missing at random ignorable

Ignorable vs. non-ignorable missing data

The likelihood of the observed data, i.e. (Y^o, R), is:

$$L(\theta, \phi | Y^o, R) = \int_{Y^m} f(Y^o, Y^m, R | \theta, \phi) dY^m = \int_{Y^m} f(Y^o, Y^m | \theta) Pr(R | Y, \phi) dY^m$$

$$\frac{d\ln L}{d\theta} = \frac{\int \left[\frac{\partial}{\partial\theta} log f(Y^o, Y^m | \theta)\right] f(Y^o, Y^m | \theta) \boxed{Pr(R|Y,\phi)} d\underline{Y^m}}{\int f(Y^o, Y^m | \theta) \boxed{Pr(R|Y,\phi)} d\underline{Y^m}}$$

$$\frac{\partial lnL}{\partial\theta} = E_{Y^m | Y^o, R}\left[U(\theta | Y^o, Y^m)\right]$$

We can note our 3 cases:

3. $Pr(R|Y,\phi) = Pr(R|Y^o, \underline{Y}^m, \phi)$ – missing not at random

non-ignorable

What do we do?

- Descriptive analysis!
- First, summarize the patterns of missingness for all variables with missing data.
  - NOTE the proportion of missingness
- Second, correlate missingness (i.e. R) with other variables
  - NOTE: There may be evidence in the data suggesting MCAR vs. MAR.
  - NOTE: You don't have any observed data to rule out MNAR!

What do we do?

- Analyze the complete-cases?
  - i.e. remove observations where there is at least missing value in the key variables.
  - This is the standard process for most statistical packages/functions
  - This can result in bias in regression coefficient estimates or in inference

MCAR: the complete cases represent a random sample of the data and there should be no bias in the estimates; however, your sample size is reduced so you lose precision to estimate the regression coefficients.

If the data are MAR (or NMAR), then the complete case analysis can result in bias for the regression coefficients.

- Example: Suppose the outcome is death and the predictors are age, sex and blood pressure. There is missing blood pressure information. Suppose the most common reason for missing blood pressure is that the participant was very close to death. Deletion of this group of very sick participants would likely bias the associations towards the null.

NOTE: In the case of *MAR*, if we can specify the correct full data likelihood, then we get unbiased (yet inefficient) estimates of measures of association. (see example next slide)

Illustration: Simulation set up

$$\underline{Y_{ij}} = \underline{\beta_1 + \beta_2 t_{ij}} + \underline{\varepsilon_{ij}}, i = 1, \ldots, m = \underline{1000}, j = \underline{1,2,3,4,5}$$

$$\text{Set the truth to be: } \beta_1 = \underline{5}, \beta_2 = \underline{0.25}$$

$$\underline{\varepsilon_{ij}} \sim N(0,\underline{1}), \underline{Cov(Y_{ij}, Y_{ik}) = \rho^{|j-k|}}$$

All subjects are observed at time 1

Subjects may then drop-out of the study at

$D_i = \underline{2,3,4,5}$ or not drop out

$$\log\left\{\frac{\Pr(D_i = \underline{k} \mid D_i \ge k, \boxed{Y_{i1}, \ldots, Y_{ik}})}{\Pr(\underline{D_i > k} \mid D_i \ge k, Y_{i1}, \ldots, Y_{ik})}\right\} = \underline{\theta_1} + \underline{\theta_2} \underbrace{Y_{ik-1}}_{\text{prior observed outcome}} + \underline{\theta_3} \underbrace{Y_{ik}}_{\text{outcome I would observe at time k if the patient doesnt dropout}}$$

→ prior observed outcome

→ outcome I would observe at time k if the patient doesnt dropout

Illustration: Simulation set up

$$Y_{ij} = \beta_1 + \beta_2 t_{ij} + \varepsilon_{ij}, i = 1,\ldots, m = 1000, j = 1,2,3,4,5$$

$$\varepsilon_{ij} \sim N(0,1), Cov(Y_{ij}, Y_{ik}) = \rho^{|j-k|}$$

$$\beta_1 = 5, \beta_2 = 0.25, \rho = 0.7$$

$$\log\left\{\frac{\Pr(D_{\mathrm{i}} = k \mid D_i \geq k, Y_{i1},\ldots, Y_{ik})}{\Pr(D_i > k \mid D_i \geq k, Y_{i1},\ldots, Y_{ik})}\right\} = \theta_1 + \theta_2 Y_{ik-1} + \theta_3 Y_{ik}$$

$$\mathrm{MCAR}: \theta_1 = -0.5, \theta_2 = \theta_3 = 0$$

$$\mathrm{MAR}: \theta_1 = -0.5, \theta_2 = 0.5, \theta_3 = 0$$

$$\mathrm{NMAR}: \theta_1 = -0.5, \theta_2 = 0, \theta_3 = 0.5$$

Illustration: Results

- Set the truth to $\beta_1 = 5, \beta_2 = 0.25$

wrong model	MCAR	MAR	NMAR
Working independence assumption (OLS/GEE with robust variance)	$\hat{\beta}_1$ 4.93 (0.04) se($\hat{\beta}_1$) $\hat{\beta}_2$ 0.26 (0.02) se($\hat{\beta}_2$)	5.02 (0.04) 0.14 (0.02) biased est β_2	5.06 (0.04) 0.10 (0.02)
WLS (AR1) correct model	4.93 (0.04) 0.25 (0.01)	4.91 (0.01) 0.27 (0.01)	4.98 (0.04) 0.20 (0.02)

NOTE: If NMAR holds, the bias in the WLS tends to be smaller than OLS/GEE and the bias decreases as the correlation among the repeated measures increases to 1.

Imputation algorithms → MAR

→ Predict or fill in y^m based on Y^o, X, or other variables

- Single conditional mean imputation

→ Assume Y_{i1} is observed for each subject

$$Y_{i2} = \beta_0 + \beta_1 Y_{i1} + \beta_2 X_{i1} + \dots + \beta_p X_{ip} + \epsilon_{ij}$$

→ fit using all observed data

Predicted / imputed $\hat{Y}_{i2}^{m} = \hat{\beta}_0 + \hat{\beta} \hat{Y}_{i1} + \hat{\beta}_2 X_{i1} + \dots + \hat{\beta}_p X_{ip}$

Y_{11}	$R_{11}=0$	Y_{11}
Y_{12}	$R_{12}=0$	Y_{12}
Y_{21}	$R_{21}=0$	Y_{21}
NA	$R_{22}=1$	$\hat{Y}_{22}$

all subjects

Imputation algorithms

- ► <u>Single predicted value imputation</u>

Y_{i1} is observed for everyone

↳ female: observed for everyone

missing Y_{i2}

⇒ using data for subjects with $R_{i2} = 0$

fit $Y_{i2} = \beta_0 + \beta_1 Y_{i1} + \beta_2 \text{female}_i + \varepsilon_{ij}$, $\varepsilon_{ij} \sim N(0, \sigma^2)$

⇒ $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2$

conditional mean

⇒ imputed value $\underbrace{Y_{i2}^{m}} = \underbrace{\left[\hat{\beta}_0 + \hat{\beta}_1 Y_{i1} + \hat{\beta}_2 \text{female}_i\right]} + \underline{\text{random draw from } N(0, \hat{\sigma}^2)}$

Imputation algorithms

- Multiple imputation = m imputed datasets

ID	Y_{i1}	Y_{i2}	$Y_{i2}^{(1)}$	$Y_{i2}^{(2)}$	…	$Y_{i2}^{(m)}$
1	5	10	10	10		10
2	10	NA	12	8		9
3	15	NA	17	22		16
4	5	12	12	12		12

Fit analysis of interest for each m

pool results

$\hat{\alpha}^{(m)}, v(\hat{\alpha}^{(m)})$

use ID 1 and 4 ⇒ fit $Y_{i2} = \beta_0 + \beta_1 Y_{i1} + \beta_2 \text{females} + \varepsilon_{ij}$

⇒ draw multiple values of Y_{i2} for ID 2 and 3

$\sum_{i=1}^{m} \hat{\alpha}^{(i)} / m$

Variance pooled $\overline{V(\hat{\alpha}^{(i)})} + \hat{Var}(\hat{\alpha}^{(m)})$

Imputation algorithms

- Matching methods

* missingness depends on age and gender

* create subsets of participants based on age and gend

	M	F
Age 40-64	y^o, y^m	
Age 65-85		
Ages 85+		

within a subset, impute y^m using a randomly selected y^o from that subset

Imputation algorithms

- Weighting Approaches: think survey sampling

→ Think of your dataset as a sampling frame

ID	Age	Y sysBP	wt
1	65	120	2
2	65	NA	
3	70	130	2
4	70	NA	
5	85	122	1
			5

$$Pr(R_i=1 \mid Age)$$

$$Pr(R_i=1 \mid Age=65) = .5$$

$$Pr(R_i=1 \mid Age=70) = .5$$

$$Pr(R_i=1 \mid Age=85) = 1$$

$$wt = \frac{1}{1-Pr(R_i=1 \mid Age)}$$

Developing an imputation model

- What should I include or not include in an imputation model for a variable with missing values, will call it the target variable below.

- The model should include all variables that are either:
  - related to the missing data mechanism
  - have distributions that differ between subjects with observed and missing values for the target variable
  - are associated with the target variable when it is not missing
  - are included in the final response model

Multiple imputation

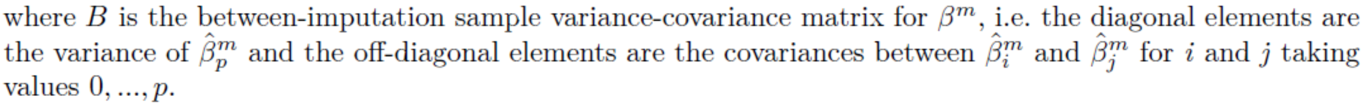
- Create M imputed datasets.

- Fit your analysis model to each of the M datasets, save $\hat{\beta}^m, \hat{V}(\hat{\beta}^m)$

3. Average your results:

$$\overline{\beta}_i = \frac{1}{M}\sum_{m=1}^{M} \hat{\beta}_i^m$$

$$\overline{V} = \frac{1}{M}\sum_{m=1}^{M} \hat{V}(\hat{\beta}^m) + \frac{M+1}{M}B$$

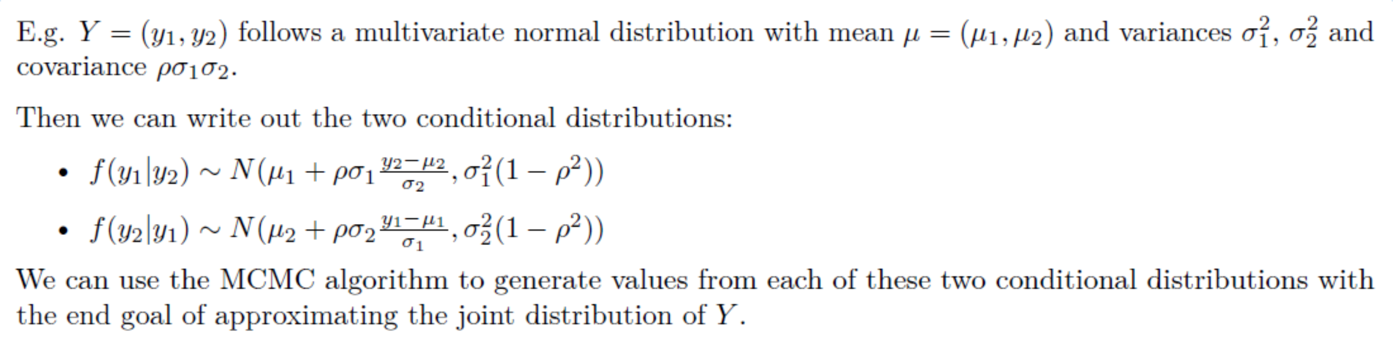


where B is the between-imputation sample variance-covariance matrix for β^m, i.e. the diagonal elements are the variance of $\hat{\beta}_p^m$ and the off-diagonal elements are the covariances between $\hat{\beta}_i^m$ and $\hat{\beta}_j^m$ for i and j taking values $0, ..., p$.

Chained equation approach

mice

- The idea here is anchored in the desire to estimate the joint distribution of a set of random variables (some values of which are missing). We may be able to derive the exact joint distribution OR we can approximate the joint distribution by deriving the set of full conditional distributions.



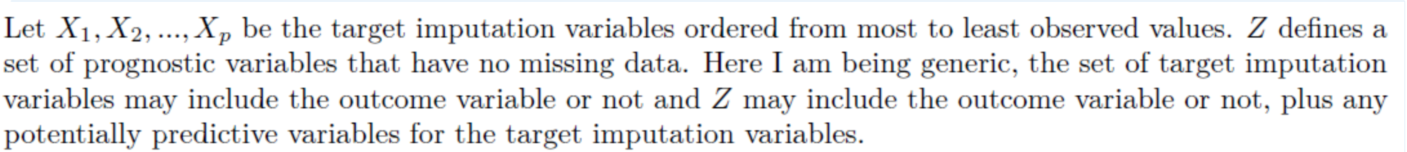
E.g. $Y = (y_1, y_2)$ follows a multivariate normal distribution with mean $\mu = (\mu_1, \mu_2)$ and variances σ_1^2, σ_2^2 and covariance $\rho\sigma_1\sigma_2$.

Then we can write out the two conditional distributions:

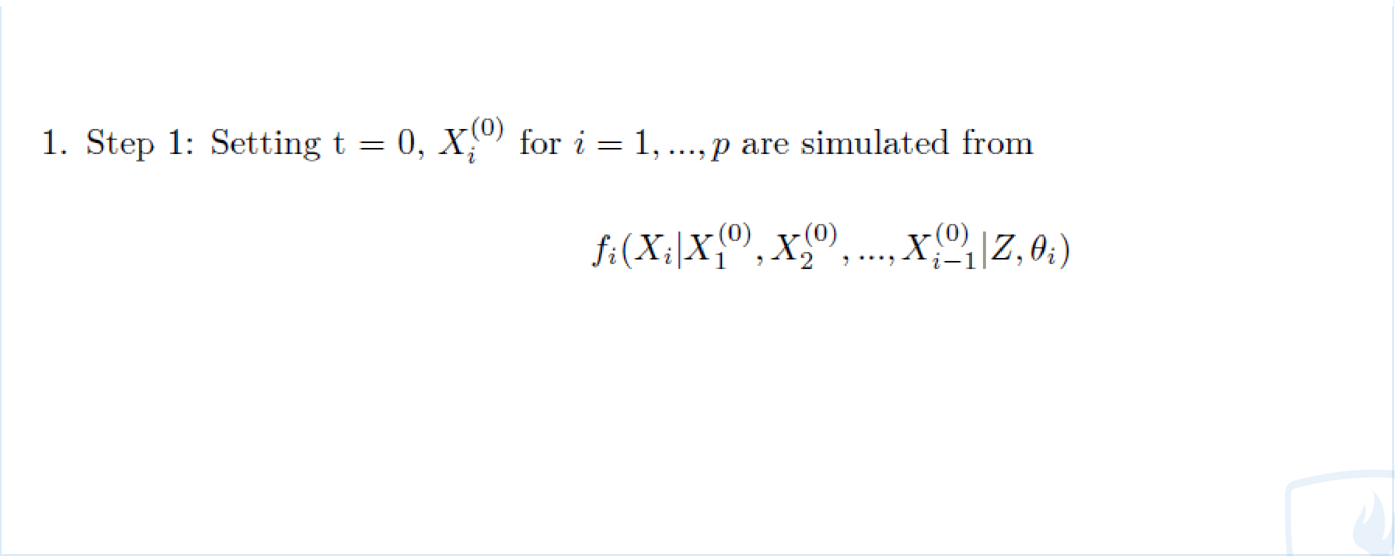
- $f(y_1|y_2) \sim N(\mu_1 + \rho\sigma_1 \frac{y_2 - \mu_2}{\sigma_2}, \sigma_1^2(1 - \rho^2))$
- $f(y_2|y_1) \sim N(\mu_2 + \rho\sigma_2 \frac{y_1 - \mu_1}{\sigma_1}, \sigma_2^2(1 - \rho^2))$

We can use the MCMC algorithm to generate values from each of these two conditional distributions with the end goal of approximating the joint distribution of Y.

Chained equation approach



Let $X_1, X_2, ..., X_p$ be the target imputation variables ordered from most to least observed values. Z defines a set of prognostic variables that have no missing data. Here I am being generic, the set of target imputation variables may include the outcome variable or not and Z may include the outcome variable or not, plus any potentially predictive variables for the target imputation variables.



1. Step 1: Setting t $= 0$, $X_i^{(0)}$ for $i = 1, ..., p$ are simulated from

$$f_i(X_i|X_1^{(0)}, X_2^{(0)}, ..., X_{i-1}^{(0)}|Z, \theta_i)$$

Chained equation approach

2. Step 2: For t = 1: obtain simulated values $X_i^{(1)}$ for $i = 1, ..., p$ from

$$g_1(X_1|X_2^{(0)}, ..., X_p^{(0)}, Z, \phi_1)$$

$$g_2(X_2|X_1^{(1)}, X_3^{(0)}, ..., X_p^{(0)}, Z, \phi_2)$$

through

$$g_p(X_p|X_1^{(1)}, X_2^{(1)}, ..., X_{p-1}^{(1)}, Z, \phi_p)$$

Then repeat this process for t = 2, . . . , b.

What is coming next …..

- On Tuesday, we will work through an example using the Nepali Anthropometry Study