# Lecture 1~~4~~5

## Missing data considerations

## R implementation of imputation approaches

# 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
  - Implement several imputation procedures using R mice package

# Missing data mechanisms

$Y = (Y^o, Y^m)$, $R$ = indicators for whether $Y$ is missing or observed

| Missing data mechanism | Definition | Ignorable? |
|---|---|---|
| Completely at random<br>$f(R \mid \theta)$<br>$f(R \mid X)$<br>↳ covariate dependent missingness | Whether or not a value is missing does not depend on observed data OR the missing value we would have observed.<br>- Think a value is missing based a coin toss with some probability of a head that does not depend on anything else | Yes<br>$\boxed{\text{Complete cases}}$ represent a random sample of original sample -> analysis of complete case will loss precision but will be unbiased |
| At random | | |
| Not at random | | |

# Missing data mechanisms

| Missing data mechanism | Definition | Ignorable? |
|---|---|---|
| Completely at random<br>$Y_i = \begin{pmatrix} Y_i^o \\ Y_i^m \end{pmatrix}$ $f(R\|X)$ $f(R\|\theta)$ | Whether or not a value is missing does not depend on observed data OR the missing value we would have observed. | Yes |
| At random<br>$f(R\|Y_i^o)$ | Whether or not a value is missing depends on observed ~~covariates~~ X, Y<br>- Think a value is missing is based on a coin toss with probability based on observed characteristics | $\boxed{\text{Yes}}$ $\boxed{f(Y\|X)}$<br>Complete cases are NOT representative of the original sample. Analysis of complete cases may be biased unless you correctly specify the model, could lose precision |
| Not at random | | |

# Missing data mechanisms

| Missing data mechanism | Definition | Ignorable? |
|---|---|---|
| Completely at random | Whether or not a value is missing does not depend on observed data OR the missing value we would have observed. | Yes |
| At random | Whether or not a value is missing depends on observed covariates | Yes |
| Not at random | Whether or not a value is missing depend on the value of the variable you would observe had it not been missing | No<br>Complete cases are a specific selection of the original sample<br>Bias! |

analysis of available data

assumption is MAR

imputation approach

$f(R \mid y^m)$


$f(y^m \mid x) \approx f(y^o \mid x)$ benchmark assumption

$= \exp(-\delta) f(y^o \mid x)$

# Imputation algorithms

- Single conditional mean imputation $Y^o = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$ $\varepsilon \sim N(0, \sigma^2)$

imputed value $\hat{y}^m = \hat\beta_0 + \hat\beta_1 X_1 + \dots + \hat\beta_p X_p$

- Single predicted value imputation

imputed value $\hat{y}^m = \hat\beta_0 + \hat\beta_1 X_1 + \dots + \hat\beta_p X_p$ + a draw from $N(0, \hat\sigma^2)$

- Multiple imputation: repeat the single predicted value imputation several times

$\hat{y}^{m(1)} \quad \dots \quad \hat{y}^{m(M)}$

- Matching methods: create subsets of individuals based on $X$: $Y^o, Y^m$

imputed values are randomly selected among $Y^o$

# Chained equation approach

$f(Y, X)$

- 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 $\sigma_1^2$, $\sigma_2^2$ 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

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

single vale prediction

$$X_1^{(0)} \Rightarrow f_1(X_1 | Z, \theta_1)$$

$$X_2^{(0)} \Rightarrow f_2\left(X_2 \middle| X_1^{(0)}, Z, \theta_2\right)$$

$$X_3^{(0)} \Rightarrow f_3\left(X_3 \middle| X_2^{(0)}, X_1^{(0)} Z, \theta_3\right)$$

...

# Chained equation approach

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

$$X_1^{(1)} \Rightarrow g_1(X_1|X_2^{(0)}, ..., X_p^{(0)}, Z, \phi_1)$$

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

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.

$X_1^{(0)} \ \dots \ X_1^{(b)}$

$X_2^{(0)} \ \dots \ X_2^{(b)}$

$\dots$

# Now, let's implement some of these approaches!