



JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Lecture 14

Missing data considerations

The material in this video is subject to the copyright of the owners of the material and is being provided for educational purposes under rules of fair use for registered students in this course only. No additional copies of the copyrighted work may be made or distributed.

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
 - ▶ 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.

- ▶ $Y_i = \begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{ini} \end{bmatrix} = \begin{bmatrix} Y^o \\ Y^m \end{bmatrix}$

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

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



Missing completely at random



Missing at random



Missing not at random



Ignorable vs. non-ignorable missing data

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

$$\begin{aligned} L(\theta, \phi | Y^o, R) &= \int_{Y^m} f(Y^o, Y^m, R | \theta, \phi) dY^m = \int_{Y^m} f(Y^o, Y^m | \theta) Pr(R | Y, \phi) dY^m \\ &= \frac{\int \left[\frac{\partial}{\partial \theta} \log f(Y^o, Y^m | \theta) \right] f(Y^o, Y^m | \theta) Pr(R | Y, \phi) dY^m}{\int f(Y^o, Y^m | \theta) Pr(R | Y, \phi) dY^m} \\ \frac{\partial \ln L}{\partial \theta} &= E_{Y^m | Y^o, R} [U(\theta | Y^o, Y^m)] \end{aligned}$$

We can note our 3 cases:

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



Ignorable vs. non-ignorable missing data

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

$$\begin{aligned} L(\theta, \phi | Y^o, R) &= \int_{Y^m} f(Y^o, Y^m, R | \theta, \phi) dY^m = \int_{Y^m} f(Y^o, Y^m | \theta) Pr(R | Y, \phi) dY^m \\ &= \frac{\int \left[\frac{\partial}{\partial \theta} \log f(Y^o, Y^m | \theta) \right] f(Y^o, Y^m | \theta) Pr(R | Y, \phi) dY^m}{\int f(Y^o, Y^m | \theta) Pr(R | Y, \phi) dY^m} \\ \frac{\partial \ln L}{\partial \theta} &= E_{Y^m | Y^o, R} [U(\theta | Y^o, Y^m)] \end{aligned}$$

We can note our 3 cases:

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



Ignorable vs. non-ignorable missing data

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

$$\begin{aligned} L(\theta, \phi | Y^o, R) &= \int_{Y^m} f(Y^o, Y^m, R | \theta, \phi) dY^m = \int_{Y^m} f(Y^o, Y^m | \theta) Pr(R | Y, \phi) dY^m \\ &= \frac{\int \left[\frac{\partial}{\partial \theta} \log f(Y^o, Y^m | \theta) \right] f(Y^o, Y^m | \theta) Pr(R | Y, \phi) dY^m}{\int f(Y^o, Y^m | \theta) Pr(R | Y, \phi) dY^m} \\ \frac{\partial \ln L}{\partial \theta} &= E_{Y^m | Y^o, R} [U(\theta | Y^o, Y^m)] \end{aligned}$$

We can note our 3 cases:

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



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

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

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

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

All subjects are observed at time 1

Subjects may then drop - out of the study at

$D_i = 2, 3, 4, 5$ or not drop out

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



Illustration: Simulation set up

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

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

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

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

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

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

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



Illustration: Results

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

	MCAR	MAR	NMAR
Working independence assumption (OLS/GEE with robust variance)	4.93 (0.04) 0.26 (0.02)	5.02 (0.04) 0.14 (0.02)	5.06 (0.04) 0.10 (0.02)
WLS (AR1)	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

- ▶ Single conditional mean imputation



Imputation algorithms

- ▶ Single predicted value imputation



Imputation algorithms

- ▶ Multiple imputation



Imputation algorithms

- ▶ Matching methods



Imputation algorithms

- ▶ Weighting Approaches: think survey sampling



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:

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

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

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 $0, \dots, p$.

Chained equation approach

- ▶ 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, \dots, 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 $i = 1, \dots, p$ are simulated from

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



Chained equation approach

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

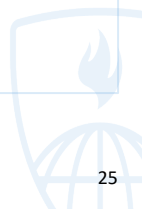
$$g_1(X_1|X_2^{(0)}, \dots, X_p^{(0)}, Z, \phi_1)$$

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

through

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

Then repeat this process for $t = 2, \dots, b$.



What is coming next

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

