Lecture 14

Elizabeth Colantuoni

3/9/2021

I. Objectives

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

II. Missing data mechanisms

Missing data is common (*VERY COMMON*) and can occur in either the outcome (Y) or the exposure variables (X). Statistical thinking about missing data is rooted in understanding/defining how the missing data was generated, i.e. the data generating mechanism for the missing data.

We will start by considering missing data within a longitudinal study to help motivate the intiution for the missing data mechanisms; specifically we will consider that subjects may drop-out of the study over time. We observe values of the outcomes up until the time of drop-out; after drop-out the outcome values are missing.

Let Y be the $n \times 1$ response vector for a given subject and subjects may drop-out after a certain portion of the study is completed. Then we can express $Y = (Y^o, Y^m)$ where Y^o are the values of Y that are observed and Y^m are the values of Y we would have observed had no values been missing. Define R to be the $n \times 1$ vector indicating which values of Y are missing (value of 1) vs. observed (value of 0). Let θ describe parameters associated with the distribution of Y and let ϕ describe parameters associated with R such that ϕ is non-overlapping with θ.

A. Missing completely at random:

Data are said to be *missing completely at random* if

$$Pr(R|Y, \phi) = Pr(R|\phi)$$

Here, the distribution of R doesn't depend on the observed (Y^o) or missing (Y^m) data.

B. Missing at random:

Data are said to be *missing at random* if

$$Pr(R|Y,\phi) = Pr(R|Y^o,\phi)$$

Here the distribution of R depends on the data Y only through Y^o, i.e. whether or not a subject will drop out is informed by the observed responses for that subject at prior times.

NOTE: *missing at random* behaves like *missing completely at random* within subsets of subjects with similar Y^o.

C. Missing not at random:

Data are said to be *missing not at random* if the distribution of R depends on Y^m,

$$Pr(R|Y,\phi) = Pr(R|Y^o,Y^m,\phi)$$

i.e. whether we get to see the outcome depends on the value of the outcome itself.

III. Ignorable vs. Non-ignorable missing data

The data generating mechanisms *missing completely at random* and *missing at random* are **ignorable**; whereas, *missing not at random* is a **non-ignorable** mechanism. If the missing data is generated by one of the two **ignorable** missing data mechanisms, the likelihood function of the observed data does not depend on the missing data mechanism.

For those of you who are interested in the more technical details, I show this below.

A. Necessary Math Facts

1. $\frac{\partial}{\partial\theta}\int f(x|\theta)dx = \int \frac{\partial}{\partial\theta} f(x|\theta)dx$ i.e. you can exchange the derivative and integral
2. $\frac{\partial}{\partial\theta} log(f(x|\theta)) = \frac{1}{f(x|\theta)}\frac{\partial}{\partial\theta} f(x|\theta) \Rightarrow \frac{\partial}{\partial\theta} f(x|\theta) = \left[\frac{\partial}{\partial\theta} log(f(x|\theta))\right] \times f(x|\theta)$
3. Combining 1. and 2.

$$\frac{\partial}{\partial\theta}\int f(x|\theta)dx = \int \frac{\partial}{\partial\theta} f(x|\theta)dx = \int \left[\frac{\partial}{\partial\theta} log(f(x|\theta))\right] \times f(x|\theta) = E_x\left[U(\theta|x)\right]$$

where $U(\theta|x)$ is the score equation for the likelihood function of θ, i.e. the derivative of the log likelihood function taken with respect to θ.

B. Derivation

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

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

Let $* = f(Y^o, Y^m|\theta)Pr(R|Y,\phi)$.

Derive the estimating equation for θ.

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

This derivation represents the Expectation-Maximization Algorithm by Dempster, Laird and Rubin (1977).

We can note our 3 cases:

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

Cases 1 and 2 are **ignorable**; $Pr(R|Y,\phi)$ cancels in both the numerator and denominator so the likelihood function for the observed data does not depend on the missing data mechanism.

Case 3 is **non-ignorable**; $Pr(R|Y,\phi)$ does not cancel! The solution for θ will be affected if the missing data is *missing not at random*.

IV: What do we do?

A. Summarize the patterns of missing data

Use exploratory data analysis to describe the patterns of missing data for each variable that contains missing values (with a special focus on your outcome and primary explanatory variables).

Describe how missingness is related to the other variables you are considering (including looking at the relationship between Y and X if X is the variable that contains missing values).

There may be evidence in the data that suggests *MCAR* over *MAR*. But be aware: you can never rule out *NMAR*.

B. Complete cases or casewise deletion

This is the standard approach for most statistical packages/procedures!

This approach can result in regression coefficients that are biased, imprecise or both!

- If the data are MCAR, then 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 of interest is death and the predictors are age, sex and blood pressure. Age and sex are recorded for each participant but blood pressure is missing. 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*, we we can specify the correct full data likelihood, then we get unbiased (yet inefficient) estimates of measures of association. See example of missingness in longitudinal studies in Lecture slides.

C. Imputation algorithms

This is a *prediction* problem, asking the question: what value of Y would I have observed if the participant responded to the question, returned to complete the survey or the measure was able to be completed?

What strategies/algorithms are available?

- Single conditional mean imputation: replace missing values with the predicted conditional expected value/mean.
- Single predicted value imputation: replace missing values using single individual predicted values. E.g. if Y contains missing values, then based on a model for Y given X, take a single value from the distribution of Y given X based on the estimated $E(Y|X)$ and $Var(Y|X)$. If normality is not assumed, draw a random residual (sampled with replacement) after the fit of the model for $Y|X$ and take estimated $E(Y|X) + residual$.
- Multiple imputation using single predicted value imputation
- Matching methods: Single or multiple predictive mean matching (PMM). Here you replace missing values with a randomly selected observed value of Y within a matched set of X or based on a measure of how "close" X from the participant with missing value Y is from participants Xs for whom we observe Y.
- Weighting methods: I won't discuss these in detail. The ideas are borrowed from the sampling literature. Treat the non-missing observations like participants that were included in your sample based on the full dataset/sampling frame. Create sampling weights such that participants with observed data represent > 1 members of the sampling frame.

D. 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

E. Multiple imputation

The idea here is that our imputation approaches rely on us assuming a model, which we estimate with error. If we complete a single imputation and fit our final analysis, then our standard error estimates account for the variance we observed in the "full dataset" (although somewhat perturbed by the imputed value) but does NOT account for the error in our model that was used to make the prediction.

Multiply imputed values allows use to evaluate the fit of our final analysis accounting for variation we observed in the data + uncertainty in our imputed values.

Repeat the imputation M times! Fit your final model to each of the M imputed datasets and obtain $\hat{\beta}^m$ and $\hat{V}(\hat{\beta}^m)$.

Our final estimates are:

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

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

where B is the between-imputation sample variance-covariance matrix for $\hat{\beta}^m$, 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 $0, ..., p$.

Choose M large enough: One rule of thumb is $100f$ where f is the fraction of cases that are incomplete.

F. 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.

1. Application to missing data

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

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

Here we are constructing sequential conditional models but where we are NOT conditioning on the full set of target variables; rather we are sequentially adding the target variables: $f_1(X_1|Z, \theta_1)$, $f_2(X_2|X_1^{(0)}, Z, \theta_2)$, $f_3(X_3|X_1^{(0)}, X_2^{(0)}, Z, \theta_3)$, ...

This process provides us with an initial starting value for each X_i.

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.

3. Step 3: Take as the imputed values the final set of $X_i^{(b)}$. Fit the analysis of interest.
4. Repeat Steps 2 and 3 until you have M imputed datasets.