Missing Data

Background

"Bayesian data analysis draws no distinction between missing data and parameters. Both are uncertain, and they have a joint posterior distribution, conditional on observed data." ~ BDA

A Bayesian model with missing data includes three parts:

1. A prior distribution for the parameters
2. A joint model for all of the data (missing and observed)
3. An inclusion model for the missingness process (if the missing data mechanism is non-ignorable)

Definitions

Let $y = (y_{obs}, y_{mis})$. Let I be the inclusion indicator. Let θ be model parameters and let ϕ be parameters governing the missing data mechanism. Then the joint distribution of of interest is

$$p(y, I|\theta, \phi) = p(y|\theta)p(I|y, \phi)$$

Definition: *inclusion indicator* – A data structure with the same size and shape as as y with 1 if the corresponding component is observed and 0 if the corresponding component is missing.

Definition: *inclusion model* – The part of the statistical model that tries to model the inclusion indicator. The nature of this model is determined by the type of missingness.

$$p(I|y_{obs}, y_{mis}, \phi)$$

Definition: *missing completely at random (MCAR)* – If the probability of missingness is the same for all observations. Cause of missingness is unrelated to the data. A simple example is random sampling. This is also called *observed at random*.

$$p(I|y_{obs}, y_{mis}, \phi) = p(I|\phi)$$

Definition: *missing at random (MAR)* – If the distribution of the missing data mechanism does not depend on the missing values. *The distribution of the missing data mechanism can depend on fully observed values in the data and parameters for the missing data mechanism.* A simple example is a stratified random sample.

$$p(I|y_{obs}, y_{mis}, \phi) = p(I|y_{obs}, \phi)$$

Definition: *ignorable missing data mechanism* – If the parameters for the missing data mechanism ϕ and the parameters for the model θ are distinct, then the missing data mechanism is said to be ignorable. (BDA frames this as ϕ and θ are independent in the prior distribution).

$$p(y_{obs}, I|\theta, \phi) = p(y_{obs}|\theta)$$

In this situation

$$p(\theta|x, y_{obs}) = p(\theta|x, y_{obs}, I)$$

Definition: *missing not at random (MNAR)* – If the missing data mechanism depends on the missing values. Neither MCAR or MAR holds. The data are missing for reasons that are unknown to us.

Creating Missing Data

We assume that there is no unit missingness. Our observations are either a simple random sample from the population or the complete population. Our approach will only focus on unit missingness, where values within an observation are missing. We will start with our complete data and then construct different obscured data sets with different types of mssingness.

Van Buuren notes that there are tests to determine MCAR vs. MAR but they aren't widely used. There are no tests to compare MAR vs. MNAR. In this case, we know the missingness for illustrative purposes.

1. MCAR

Randomly drop values.

$$p(I|\phi)$$

In this case, ϕ will just be a probability.

2. MAR with ignorable missing data mechanism

Drop values conditional on variables that will be included in the final model of interest. Furthermore, the prior distribution for θ and ϕ need to be independent.

$$p(I|y_{obs}, y_{mis}, \phi) = p(I|y_{obs}, \phi)$$

3. MAR with non-ignorable missing data mechanism

Drop values conditional on variables that will not be included in the final model of interest. In this case, we will need to use a variable that is not a covariate in the regression model to model the missingness.

Van Buuren lists two remedial measures for non-ignorability:

1. Expand the data (in a situation where high weight teens are reluctant to report weights, use waist circumference to model I)
2. Create separate models for $I = 0$ and $I = 1$.

   In general, the formulation of nonignorable models should be driven by knowledge about the process that created the missing data. Any such methods need to be explained and justified as part of the statistical analysis.

We will employ option 1.

4. MNAR

Drop values conditional on data we observe but then drop from the data set.

https://stefvanbuuren.name/fimd/sec-idconcepts.html