On Missing Data

Drafted by Haoda Fu, Ph.D.

Statistics for Data Scientists Series

October 17, 2018

Advanced Analytics and Data Sciences (AADS)
Eli Lilly and Company

Why

Why this topic is important

Missing data are common problems for data scientists.

Why

Why this topic is important

Missing data are common problems for data scientists.

Connections

- Missing data and causal inference are closed connected and both are important to our digital health.
- Different types of missing mechanism: missing completely at random, missing at random, and missing not at random.
- Different methods to handle missing data.

Notation

- $Y = (Y_{ij})$: complete data matrix.
- $M = (M_{ij})$: missing-data indicator matrix.
- ϕ: unknown parameter.
- Y_{obs} and Y_{mis}: the observed and missing components of Y.
- $f(\cdot)$: probability density function.

Missing Completely at Random (MCAR)

Definition (MCAR)

MCAR is defined that the M does not depend on Y,

$$f(M|Y,\phi) = f(M|\phi), \quad \forall Y, \phi.$$

Note: this assumption does not mean that the pattern itself is random, but rather that the missingness does not depend on the data values.

Missing at Random (MAR)

Definition (MAR)

MAR is defined that the M only depends on Y_{obs},

$$f(M|Y,\phi) = f(M|Y_{obs},\phi), \quad \forall Y_{obs},\phi.$$

Note: This assumption is weaker than the MCAR and it is the most widely used assumption for clinical trials.

Missing Not at Random (MNAR)

Definition (MNAR)

The mechanism is called MNAR if the distribution of M depends on the Y_{miss}.

Note: Assumptions are not verifiable. It is often used for robustness evaluation. Methods handle MNAR include pattern mixture models and selection models.

Parametric Likelihood-Based Analysis Under MAR

Let θ be the interest parameter, i.e. $f(Y|\theta) = f(Y_{obs}, Y_{mis}|\theta)$, and ψ be the parameter for missing mechanisms.

$$
\begin{aligned}
f(Y, M|\theta, \psi) &= f(Y|\theta) f(M|Y, \psi) \\
f(y_{obs}, m|\theta, \psi) &= \int f(y_{obs}, Y_{mis}|\theta) f(m|y_{obs}, Y_{mis}, \psi) dY_{mis} \\
&= f(m|Y_{obs}, \psi) \int f(y_{obs}, Y_{mis}|\theta) dY_{mis} \\
&\triangleq f(M|Y_{obs}, \psi) f(Y_{obs}|\theta)
\end{aligned}
$$

Remarks: the lower case means the observed value. The dY_{mis} can only be calculated when M is observed. Also, $f(Y_{obs}|\theta)$ is not the same pdf as if we only observed iid data, so that we use $\triangleq$ notation instead of regular equal to, because we define $f(Y_{obs}|\theta) = \int f(y_{obs}, Y_{mis}|\theta) dY_{mis}$.