Missing Data Mechanisms

We formalize the concepts of missing-data mechanisms by following the notation in Little and Rubin (2002).

Notation in missing data mechanisms

- Data: $\boldsymbol{Y} = (y_{ij})_{n \times K}$ where ith row $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{iK})$, y_{ij} is the value of variable $\boldsymbol{Y_j}$ for observation i, and $\boldsymbol{Y}$ is characterized by $\boldsymbol{\theta}$
- Missing data indicator matrix: $\boldsymbol{M} = (m_{ij})$, $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is present; and $\boldsymbol{M}$ is characterized by unknown parameter $\boldsymbol{\psi}$.
- The missing-data mechanism is characterized in terms of the conditional distribution of $\boldsymbol{M}$ given $\boldsymbol{Y}$, $f(\boldsymbol{M}|\boldsymbol{Y}, \boldsymbol{\psi})$.

- **Missing completely at random (MCAR)**
  This means there is no relationship between the missingness of the data and any values, observed or missing. Those missing data points are a random subset of the data.

$$f(\boldsymbol{M}|\boldsymbol{Y}, \boldsymbol{\psi}) = f(\boldsymbol{M}|\boldsymbol{\psi}) \text{ for all } \boldsymbol{Y}, \boldsymbol{\psi}.$$

Missing Data Mechanisms

- **Missing at random (MAR)**
  Let $\boldsymbol{Y} = (\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{mis})$ where $\boldsymbol{Y}_{obs}$ is defined as the observed component of $\boldsymbol{Y}$ and $\boldsymbol{Y}_{mis}$ is the missing component. When the missingness depends on $\boldsymbol{Y}_{obs}$, the missing-data mechanism is said to be MAR if

$$f(\boldsymbol{M}|\boldsymbol{Y}, \psi) = f(\boldsymbol{M}|\boldsymbol{Y}_{obs}, \psi) \text{ for all } \boldsymbol{Y}_{mis}, \psi.$$

  Whether an observation is missing has nothing to do with the missing values, but it does have to do with the values of observed variables.

- **Missing not at random (MNAR)**
  Under MNAR, the distribution of $\boldsymbol{M}$ depends the data $\boldsymbol{Y}$. This means there is a relationship between the propensity of a value to be missing and its values. For instance, people with high incomes tend to not providing their salaries in surveys.

Methods on Handling Missing Data

- **Complete case analysis:** the complete-case analysis only analyzes the complete observations in a dataset. It is easy to implement but only valid under MCAR.
- **Single imputation**
  - **Mean imputation:** mean calculated based on the complete observations in the dataset.
  - **Regression imputation:** predictions from the regression model built on complete observations.
  - **Stochastic regression imputation:** predictions from the regression model **plus** a random draw from the estimated distribution of residuals.
  - **Logistic regression imputation:** predictions from the logistic regression model built on complete observations.
  - **"Worst-rank" method:** Lachin (1999) proposed "worst-rank" analysis by assigning more extreme values (values indicating "worst" treatment effects) than observed values as the imputed values for missing data. All missing values share the same values (ranks) if a worst-rank analysis applies.
  - **"Best-worst and worst-best" method:** suppose the treatment A is the beneficial group and B is the placebo. If the "best - worst" method is adopted, the imputed values for missing responses in treatment A will be values representing harmful outcomes (i.e., the "worst" values among observed values). While in treatment B, the imputed values will be values representing beneficial outcomes (i.e., the "best" values among observed values).

Maximum Likelihood Method

- **Direct maximum likelihood method** is a method handling missing data without imputing missing values, and it is valid under MAR assumption.
- Under MAR, we have

$$f(\boldsymbol{Y}_{obs}, \boldsymbol{M}|\boldsymbol{\theta}, \boldsymbol{\psi}) = f(\boldsymbol{M}|\boldsymbol{Y}_{obs}, \boldsymbol{\psi}) f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta}),$$

where we are interested in the inference for $\boldsymbol{\theta}$. If $\boldsymbol{\theta}, \boldsymbol{\psi}$ are distinct, as the distribution $f(\boldsymbol{M}|\boldsymbol{Y}_{obs}, \boldsymbol{\psi})$ does not depend on $\boldsymbol{\theta}$, to estimate the ML of $\boldsymbol{\theta}$ is equivalent to maximize the likelihood $f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta})$; i.e, to ignore the missing data.

Maximum Likelihood Method with Regression Model

- Consider the linear regression model with the following format $\boldsymbol{y} = \tilde{\boldsymbol{X}}^{\mathsf{T}}\beta + \boldsymbol{\epsilon}$
- $\tilde{\boldsymbol{X}} = (1, \boldsymbol{X}^{\mathsf{T}})^{\mathsf{T}}$ and the predictors $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)^{\mathsf{T}} \sim N_p(\boldsymbol{\mu_X}, \Sigma_{\boldsymbol{X}})$
- $\epsilon_i \sim N(0, \sigma^2)$.
- Some values are missing within n independent observations $(y_i, \boldsymbol{X}_i^{\mathsf{T}})^{\mathsf{T}}$, $i = 1, \ldots, n$.
- Based on the normality and independence assumptions of $\boldsymbol{y}$ and $\boldsymbol{X}$, we have

$$(\boldsymbol{y}, \boldsymbol{X}) \sim N(\mu_{\boldsymbol{y},\boldsymbol{X}}, \Sigma_{\boldsymbol{y},\boldsymbol{X}})$$

where $\boldsymbol{\mu_{y,X}} = \begin{pmatrix} \boldsymbol{\mu_y} \\ \boldsymbol{\mu_X} \end{pmatrix}$ and $\Sigma_{\boldsymbol{y},\boldsymbol{X}} = \begin{pmatrix} \Sigma_{\boldsymbol{y}} & \Sigma_{\boldsymbol{y},\boldsymbol{X}} \\ \Sigma_{\boldsymbol{X},\boldsymbol{y}} & \Sigma_{\boldsymbol{X}} \end{pmatrix}$.

- The parameters $\boldsymbol{\theta} = (\boldsymbol{\mu_{y,X}}, \Sigma_{\boldsymbol{y},\boldsymbol{X}})$ are estimated through the expectation maximization algorithm (EM) algorithm.

Maximum Likelihood Method with Regression Model

Steps of EM algorithm

- **Initiate** $\theta^{(0)}$; $\theta^{(t)}$ is the estimate of θ at the tth iteration.
- **E step:** compute the expectation of complete-data log-likelihood with respect to the conditional distribution of $\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs}$ with $\theta^{(t)}$, i.e.:

$$Q(\theta|\theta^{(t)}) = E[l(\boldsymbol{Y};\theta)|\boldsymbol{Y}_{obs};\theta^{(t)}] = \int l(Y;\theta) f(\boldsymbol{Y}_{mis}|\boldsymbol{Y}_{obs};\theta^{(t)}) d\boldsymbol{Y}_{mis}.$$

- **M step:** maximize the Q function to obtain $\theta^{(t+1)}$:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}).$$

- Iterate between E step and M step until the change in function Q is very small.

After the θ has been estimated by EM algorithm, we can have the estimate of coefficient $\boldsymbol{\beta} = (\boldsymbol{\mu_y} - \Sigma_{\boldsymbol{y},\boldsymbol{X}}\Sigma_{\boldsymbol{X}}^{-1}\boldsymbol{\mu_X}, \Sigma_{\boldsymbol{y},\boldsymbol{X}}\Sigma_{\boldsymbol{X}}^{-1})^{\intercal}$. Also, the standard deviations can be estimated by

$$\mathbb{V}[\boldsymbol{\beta}] = \text{diag}(\boldsymbol{C}), \quad \text{with}$$

$$\boldsymbol{C} = (\Sigma_{\boldsymbol{y}} - \boldsymbol{\beta}^{\top}\Sigma_{\boldsymbol{X}}\boldsymbol{\beta}) \left(\left(0_{p+1}, (0_p, \Sigma_{\boldsymbol{X}})^{\top}\right)^{\top} + \left(1, \boldsymbol{\mu_X}^{\top}\right)^{\top} \left(1, \boldsymbol{\mu_X}^{\top}\right) \right)^{-1} / n.$$

Multiple Imputation

Multiple imputation: an alternative method for dealing with missing data under MAR.

Steps of multiple imputation:

- **Imputation step:** m imputations are conducted. Hence, m completed datasets are generated by replacing the missing values with the imputed values m times. Usually, setting $m = 50$ or higher is acceptable to reduce the sampling uncertainty from the imputation process.
- **The complete-data analysis step:** a desirable statistical analysis is conducted individually on each complete dataset generated from the previous step.
- **Pooling step:** collect m statistical inference results (e.g., parameter estimates and their standard errors) from the previous step. Based on the Rubin (2004)'s rules, overall parameter estimates and their standard errors, confidence intervals and p-values can be generated by combining separate results.