Multiple Imputation using Chained Rules

0. Reference

- Multiple imputation using chained equations: Issues and guidance for practice, Ian R.White et al.
- Multiple Imputation After 18+ Years, Rubin.
- Multiple Imputation for Nonresponse in Surveys, Rubin
- 연세대학교 임종호 교수님의 missing data analysis 강의

1. About Multiple Imputation

불완전 데이터 분야를 공부하다보면 루빈 통계학자의 이름이 정말 많이 나온다. EM의 아이디어도 Rubin이 제안했고 오늘 살펴볼 MI(Multiple Imputation) 또한 Rubin의 아이디어에서 시작했기 때문이다. Rubin은 1987년 그의 저서에서 MI의 개념을 아래와 같이 제안하였다.

1. Generate m realization realization of $\mathbf{y}_{mis}^{*(j)},\ j = 1, 2, \cdots, m$ from the posterior predictive distribution,

$$\mathbf{y}_{mis}^{*(j)} \sim P(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \boldsymbol{\delta}) = \int P(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \boldsymbol{\delta}; \eta) P(\eta \mid \mathbf{y}_{obs}, \boldsymbol{\delta}) d\eta \tag{1}$$

(1)은 $\mathbf{y}_{mis}$을 generate하기 위해서 베이지안 관점에서 접근하여, $\mathbf{y}_{mis}$의 사후분포를 유도하였음을 알 수 있다. 빈도론자 입장에서는 아래와 같이 $\mathbf{y}_{mis}$를 generate한다.

$$\mathbf{y}_{mis}^{*(j)} \sim P(\mathbf{y}_{mis} \mid \mathbf{y}_{obs}, \boldsymbol{\delta}; \hat{\eta}) \tag{2}$$

(2)에서는 missing mechanism으로 인해 발생하는 모수인 η가 mle이든, 어떤 추정량에 의해서 추정이 되어서 그 값을 조건으로 한 분포에서 $\mathbf{y}_{mis}$를 generate함을 알 수 있다. 현재 Imputation 분야에서는 90%~95%가 (1)의 베이지안 방법을 사용하고 나머지가 (2)의 빈도론자 방법을 사용한다고 한다.

주목해야할 점은, $\mathbf{y}_{mis}^{*(j)}$가 scalar 값이 아니라 벡터라는 것이다. 즉, $\mathbf{y}_{mis}^{*(j)}$가 의미하는 바는, (1)의 사후분포에서 표본을 뽑는 과정을 총 m번 반복하고 뽑을 때마다 1개만 뽑는 것이 아니라 결측치 갯수만큼 뽑는 것이다. MI는 결측치 갯수가 여러개일때 사용하는 방법으로, 각 iteration 별로 표본을 여러개의 결측치 갯수만큼 뽑는다.

2. Obtain $(\hat{\eta}^{*(j)}, \hat{V}^{*(j)})$ using $\mathbf{y}_{com}^{*(j)} = (\mathbf{y}_{obs}, \mathbf{y}_{mis}^{*(j)}),\ j = 1, 2, \cdots, m$

1번의 논의를 연장하여 $\mathbf{y}_{com}^{*(j)}$가 의미하는 바를 생각해보면 j번째 표본을 뽑았을 때의 관측된 y들과 결측치에 대한 표본이다. 따라서 $\mathbf{y}_{com}^{*(j)}$도 표본을 뽑은 iteration만큼 생기는 것이다. 따라서 $\mathbf{y}_{com}^{*(j)}$를 이용하여 $(\hat{\eta}^{*(j)}, \hat{V}^{*(j)})$를 유도하는 과정도 표본을 뽑은 만큼 한다. 여기서 $\hat{V}^{*(j)}$는 $\hat{\eta}^{*(j)}$의 variance estimator이다.

3. Combine m estimates:

   (a) Point estimator of η:

$$\hat{\eta}_{mi} = m^{-1} \sum_{j=1}^{n} \hat{\eta}_I^{*(j)} \tag{3}$$

   여기서 $\hat{\eta}_I^{*(j)} = \hat{\eta}(\mathbf{y}_{com}^{*(j)})$ 인데, 즉 complete sample을 이용한 η에 대한 추정량이라는 뜻이다.

   (b) Variance estimator of $\hat{\eta}_{MI}$:

$$\hat{V}_{MI} = W_m + (1 + m^{-1}) B_m \tag{4}$$

   where $W_m = m^{-1} \sum_{t=1}^{m} \hat{V}_n^{*(j)}$ and $B_m = (m-1)^{-1} \sum_{j=1}^{m} \left(\hat{\eta}_I^{*(j)} - \hat{\eta}_{mi} \right)^2$
   W_m는 within variance을, B_m은 Between variance을 의미하고 이들의 선형결합으로 $\hat{\eta}_{MI}$의 분산을 추정한다는 뜻이다.

4. 위 과정은 모두 MAR 가정 하에서 진행한다.

MICE는 MI를 이용한 방법이므로, 크게 위 1번~4번 과정을 거친다고 생각하면 된다. 그렇다면, Rubin은 어떻게 (3)과 (4)를 도출하였을까? 이에 대한 bayesian justification을 아래와 같이 참고사항으로 정리해보았다.

Bayesian Justification for the Rubin's Derivation

(3)에 대한 justification을 위해, 아래와 같은 Monte Carlo Approximation을 생각하자.

$$\lim_{m \to \infty} m^{-1} \sum_{j=1}^{m} \hat{\eta}(\mathbf{y}_{com}^{*(j)}) = E\left\{ \hat{\eta}(\mathbf{y}_{com}) \mid \mathbf{y}_{obs}, \delta \right\} \tag{5}$$

(5)와 같이 쓸 수 있는 이유는, $\mathbf{y}_{com}^{*(j)} = (\mathbf{y}_{obs}, \mathbf{y}_{mis}^{*(j)})$가 $f(\mathbf{y}_{com} \mid \mathbf{y}_{obs}, \delta)$로부터 독립적으로 생성되었다고 가정하기 때문이다. 또한 (5)에서 $\mathbf{y}_{com}^{*(j)}$는 realizations of $\mathbf{y}_{com}$이다. 즉, (5)의 우변에 있는 $\hat{\eta}(\mathbf{y}_{com})$은 확률변수 $\mathbf{y}_{com}$의 function이다. 이를 $\hat{\eta}(\mathbf{y}_{com}) = E(\eta \mid \mathbf{y}_{com})$으로 둔다면,

$$\begin{aligned} \lim_{m \to \infty} m^{-1} \sum_{j=1}^{m} \hat{\eta}(\mathbf{y}_{com}^{*(j)}) &= E\left\{ E(\eta \mid \mathbf{y}_{com}) \mid \mathbf{y}_{obs}, \delta \right\} \\ &= E\left\{ \eta \mid \mathbf{y}_{obs}, \delta \right\} \end{aligned} \tag{6}$$

(6)으로부터 $\hat{\eta}(\mathbf{y}_{com}^{*(j)})$의 Monte Carlo Approximation은 $E\{\eta \mid \mathbf{y}_{obs}, \delta\}$ 즉, posterior epectation과 일치한다. 따라서 $\hat{\eta}(\mathbf{y}_{com}^{*(j)})$ 을 이용한 (3)에서의 MI 추정량도 이와 일치한다.
(4)에 대한 justification을 위해, $\hat{V}(\mathbf{y}_{com}) = V(\eta \mid \mathbf{y}_{com})$, $\hat{\psi}(\mathbf{y}_{com}) = E\{\eta \mid \mathbf{y}_{com}\}$을 이용하여 아래와 같은 Monte Carlo Approximation을 생각하자.

$$\lim_{m \to \infty} W_m = \lim_{m \to \infty} m^{-1} \sum_{t=1}^{m} \hat{V}_n^{*(j)} = E\left\{ \hat{V}(\mathbf{y}_{com}) \mid \mathbf{y}_{obs}, \delta \right\} = E\left\{ V(\eta \mid \mathbf{y}_{com}) \mid \mathbf{y}_{obs}, \delta \right\} \tag{7}$$

$$\lim_{m\to\infty} B_m = \lim_{m\to\infty} (m-1)^{-1} \sum_{j=1}^{m} \left(\hat{\eta}_I^{*(j)} - \hat{\eta}_{mi}\right)^2 = V\left\{\hat{\psi}(\mathbf{y}_{com}) \mid \mathbf{y}_{obs}, \delta\right\} = V\left\{E\left\{\eta \mid \mathbf{y}_{com}\right\} \mid \mathbf{y}_{obs}, \delta\right\} \tag{8}$$

(7)과 (8)을 합치면 아래와 같다.

$$\begin{aligned}\lim_{m\to\infty} \hat{V}_{MI} &= \lim_{m\to\infty} W_m + \lim_{m\to\infty} B_m \\ &= E\left\{V(\eta \mid \mathbf{y}_{com}) \mid \mathbf{y}_{obs}, \delta\right\} + V\left\{E\left\{\eta \mid \mathbf{y}_{com}\right\} \mid \mathbf{y}_{obs}, \delta\right\} \\ &= V(\eta \mid \mathbf{y}_{obs}, \delta)\end{aligned} \tag{9}$$

(9)로부터 Rubin의 variance estimator는 posterior variance의 approximation과 일치한다.
결국, (6)과 (9)로부터, Rubin이 제안한 추정량과 분산 추정량은 사후 분포의 기댓값과 분산의 approximation과 일치하기 때문에 justify가 되는 것이다.

2. MICE: Imputing different variable types

MICE의 중요한 특징 중 하나는 여러 종류의 변수 타입을 다룰 수 있다는 것이다 (continuous, binary, unordered categorical, ordered categorical) 여기서부터는 z를 다음과 같이 정의한다: 결측치가 없는, 다른 완전한 변수들인 $\mathbf{x} = (x_1, \cdots, x_k)'$와의 관계를 이용하여 결측치를 impute하고 싶은 대상 변수.$\mathbf{x}$를 절편을 포함한 변수로, n_{obs}를 관찰된 z 값이라고 생각하자.

2.1 Continuous Variables

연속형 변수에 대해서 정규분포 가정을 하고 회귀 모형을 적합한다. 정규분포를 따르지 않는 skewed 된 분포는 추후에 transformation 파트에서 다룬다.

$$z \mid \mathbf{x}; \boldsymbol{\beta} \sim N(\boldsymbol{\beta}\mathbf{x}, \sigma^2) \tag{10}$$

$\widehat{\boldsymbol{\beta}}$은 길이가 k인 추정량 행벡터이고 $\mathbf{x}$는 길이가 k인 열벡터이다. 즉, (10)의 회귀 모형은 관찰된 z에 대해서 적합하는 것으로 논문에서는 (10)으로 썼지만 아래 첨자를 쓰면 마치 $y_i \mid x_i$ 회귀 모형처럼 볼 수 있을 것 같다.
이제 $\mathbf{V}$를 $\widehat{\boldsymbol{\beta}}$의 estimated covariance matrix, $\hat{\sigma}$를 estimated root mean-squared error라고 하자. 이제 imputation parameters $\sigma^*, \boldsymbol{\beta}^*$를 $\sigma, \boldsymbol{\beta}$의 exact joint posterior distribution에서 뽑는다.[1]

$$\sigma^* = \hat{\sigma}\sqrt{(n_{obs} - k)/g}$$

$$\boldsymbol{\beta}^* = \widehat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}} \boldsymbol{u}_1 \mathbf{V}^{1/2}$$

이제 z_i의 결측치에 대한 imputed values z_i^*를 아래와 같이 얻는다.

$$z_i^* = \boldsymbol{\beta}^* \mathbf{x}_i + u_{2i}\sigma^*$$

[1]

- Multiple Imputation for Nonresponse in Surveys, Rubin

notation을 정리하면 아래와 같다.

- g: random draw from a χ^2 distribution on $n_{obs} - k$ degrees of freedom
- $\boldsymbol{u}_1$: row vector of k independent random draws from a standard normal distribution
- $\mathbf{V}^{1/2}$: cholesky decomposition of $\mathbf{V}$
- u_{2i}: random draw from a standard normal distribution

2.2 Binary Variables

binary z를 결측치가 없는 완전한 변수 $\mathbf{x}$로부터 impute하기 위해서는 로지스틱 모형이 사용된다.

$$logit\ P(z = 1 \mid \mathbf{x}; \boldsymbol{\beta}) = \boldsymbol{\beta}\mathbf{x} \tag{11}$$

아래와 같이 notation을 정의한다.

- $\widehat{\boldsymbol{\beta}}$: estimated parameter from fitting this model to individuals with the observed z
- $\mathbf{V}$: estimated variance-covariance matrix
- $\boldsymbol{\beta}^*$: a draw from the posterior distribution of $\boldsymbol{\beta}$, approximated by MVN$(\widehat{\boldsymbol{\beta}}, \mathbf{V})$
- For each missing observation z_i, $p_i^* = [1 + exp\,(-\boldsymbol{\beta}^* \mathbf{x}_i)]^{-1}$
- u_i: a random draw from a uniform distribution on (0, 1)

이제 imputed value, z_i^*를 아래와 같이 뽑는다.

$$z_i^* = \begin{cases} 1 & if\ u_i < p_i^* \\ 0 & o.w \end{cases}$$

2.3 Unordered categorical variables

multicategory 변수에 대해서는 범주를 3개 이상으로 확장한 multicategory 로지스틱 모형을 적용한다. 총 L개의 클래스가 있다고 가정한다.

$$P(z = l \mid \mathbf{x}; \boldsymbol{\beta}) = \left[\sum_{l'=1}^{L} exp(\boldsymbol{\beta}_{l'} \mathbf{x}) \right]^{-1} exp(\boldsymbol{\beta}_l \mathbf{x}) \tag{12}$$

결측치 z_i에 대한 imputed value인 z_i^*는 아래와 같이 정의한다.

$$z_i^* = 1 + \sum_{l=1}^{L-1} I(u_i > c_{il})$$

이제 notation을 정의하자.

- $\boldsymbol{\beta}_l$: vector of dimension $k = dim(\mathbf{x})$ and $\boldsymbol{\beta}_1 = \mathbf{0}$. 여기서 첫번째 원소를 0으로 두는 이유는 baseline을 정하기 위함이다.
- $\boldsymbol{\beta}^*$: usual random draw from normal approximation to the posterior distirbution of $\boldsymbol{\beta} = (\boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_L)$, a vector of length $k(L-1)$
- $p^*_{il} = P(z_i = l \mid \mathbf{x}_i; \boldsymbol{\beta}^*)$: drawn class membership probabilities
- $c_{il} = \sum_{l'=1}^{l} p^*_{il'}$
- u_i : random draw from a uniform distribution on $(0, 1)$

2.4 Ordered Categorical Variables

클래스의 갯수가 $L > 2$개인 ordered categorical variables z에 대한 imputation은 multinomial 로지스틱 모형이나 proportional odds 모형을 이용한다. 여기서는 proportional odds 모형을 사용하는 방법에 대해서 알아본다. 2.3에서 살펴본 multinomial 로지스틱 모형은 클래스별로 다른 linear predictor을 가정하였다. (12)에서 기울기 $\boldsymbol{\beta}_l$에 첨자 l이 있는 것을 통해 알 수 있다. proportional odds 모형은 하나의 linear predictor을 아래와 같이 가정한다.

$$logit\ P(z \leq l \mid \mathbf{x}; \boldsymbol{\beta}, \zeta) = \zeta_l - \boldsymbol{\beta}\mathbf{x} \tag{13}$$

여기서 $\zeta_0 = -\infty < \zeta_1 < \cdots < \zeta_L = \infty$, $\zeta = (\zeta_1, \cdots, \zeta_{L-1})$이고 클래스별로 다른 절편을 가정한다는 뜻이다. linear predictor인 $\boldsymbol{\beta}$아래 첨자가 없으므로 클래스별로 같은 기울기를 가정한다는 뜻이다. 그 이후는 2.3과 동일하다.

$$z^*_i = 1 + \sum_{l=1}^{L-1} I(u_i > c_{il})$$

아래와 같이 notation을 정의하자.

- $\boldsymbol{\beta}^*, \zeta^*$: random draw from a normal approximation to their posterior distirubution.
- $p^*_{il} = P(z_i \leq l \mid \mathbf{x}_i; \boldsymbol{\beta}^*, \zeta^*) - P(z_i \leq l-1 \mid \mathbf{x}_i; \boldsymbol{\beta}^*, \zeta^*)$: estimated probability of individual i belonging to class $l = 1, \cdots, L$
- $c_{il} = \sum_{l'=1}^{l} p^*_{il'}$
- u_i : random draw from a uniform distribution on $(0, 1)$