Missing Data: Theory & Methods

Chapter 5

Multiple Imputation & Bayesian Analysis

Contents

The Bayesian Paradigm

- In order to adequately discuss multiple imputation and fully Bayesidan methods for missing data, we need to discuss the basic aspects of the Bayesian paradigm.
- As always, let Y denote the full data, and let its density be denoted by $p_Y(y|\theta)$, where θ is the parameter of interest. So the likelihood for a random sample $D \equiv (Y_1, \cdots, Y_n)$ is

$$L(D|\theta) = \prod_{i=1}^{n} p_Y(Y_i|\theta).$$

- In the frequentist mindset, we think of θ as having a true value θ_0 which is unknown but fixed.
- In the Bayesian mind-set, θ is treated not as fixed but random. Specifically, we express our uncertainty about θ by specifying a prior distribution for it. Denote the prior density of θ by $q(\theta)$.

The Bayesian Paradigm

- The word "prior" is used to emphasize the fact that it is the density of θ before the data $D \equiv (Y_1, \cdots, Y_n)$ are observed.
- By Bayes theorem, we can derive the distribution of θ after incorporating information in the data D, that is,

$$p(\theta|D) = \frac{L(D|\theta)q(\theta)}{\int L(D|\theta')q(\theta')d\theta'},$$

  which is called the posterior distribution of θ given D.
- The denominator, $L(D) \equiv \int L(D|\theta')q(\theta')d\theta'$, is the normalizing constant of the posterior distribution of θ. It is also the marginal distribution of D.
- For most inference problems, $p(D)$ does not have a closed form.
- Bayesian inference about θ is primarily based on the posterior distribution of θ.

The Bayesian Paradigm

- For example, one can compute various posterior summaries, such as the mean, median, mode, variance, and quantiles. For example, the posterior mean of θ is given by

$$E[\theta|D] = \int \theta p(\theta|D) d\theta,$$

  and the posterior mode is given by $\theta^m \equiv \arg\max_\theta p(\theta|D)$.
- Often, the prior $q(\theta)$ is specified using a parametric family of distributions such that

$$q(\theta) = q(\theta|\gamma),$$

  where γ is called the hyperparameter.

The Bayesian Paradigm

Bayesian procedure viewed in the frequentist paradigm

- What would be the situation if we evaluate the Bayesian estimator, say the posterior mode θ^m, under the frequentist paradigm (e.g., in terms of consistency to and concentration around a true value θ_0)?
- Note that

$$\theta^m = \arg\max_\theta \frac{L(D|\theta)q(\theta)}{L(D)} = \arg\max_\theta L(D|\theta)q(\theta),$$

  and that as $n \to \infty$, $L(D|\theta)$ becomes more and more curved with respect to θ (the quadrature of log-likelihood is sum of iid terms).
- So asymptotically, the curvature of $p(\theta|D)$ is going to be determined by $L(D|\theta)$ so that $\theta^m \approx \arg\max_\theta L(D|\theta)$, the MLE of θ.
- Hence, θ^m is consistent and efficient (when viewed as a frequentist estimator for θ_0).

Examples of deriving posterior distribution

- Suppose $Y_1, \cdots, Y_n$ are iid Bin(1, θ) given θ and that, as the prior, $\theta \sim$ Beta(α, λ), where α and λ are hyperparameters.
- That is,

$$L(D|\theta) = \theta^{\sum_{i=1}^n Y_i}(1-\theta)^{n-\sum_{i=1}^n Y_i},$$

and

$$q(\theta) = \frac{\Gamma(\alpha+\lambda)}{\Gamma(\alpha)\Gamma(\lambda)}\theta^{\alpha-1}(1-\theta)^{\lambda-1}.$$

- So the kernel of $p(\theta|D)$ can be written as

$$p(\theta|D) \propto L(D|\theta)q(\theta) \propto \theta^{\sum_{i=1}^n Y_i+\alpha-1}(1-\theta)^{n-\sum_{i=1}^n Y_i+\lambda-1}.$$

Examples of deriving posterior distribution

- We recognize this kernel as that of Beta$(\sum_{i=1}^n Y_i + \alpha, n - \sum_{i=1}^n Y_i + \lambda)$. So

$$\theta|D \sim \text{Beta}\left(\sum_{i=1}^n Y_i + \alpha, n - \sum_{i=1}^n Y_i + \lambda\right),$$

and

$$p(\theta|D) = \frac{\Gamma(\alpha+\lambda+n)}{\Gamma(\sum_{i=1}^n Y_i + \alpha)\Gamma(n - \sum_{i=1}^n Y_i + \lambda)} \times \theta^{\sum_{i=1}^n Y_i + \alpha - 1}(1-\theta)^{n - \sum_{i=1}^n Y_i + \lambda - 1}.$$

- In deriving posterior densities, an often used technique is to try and recognize the kernel of the posterior density of θ.
- This avoids a direct computation of $L(D)$. This technique saves lots of time in the derivation.

Examples of deriving posterior distribution

- Suppose $Y_1, \cdots, Y_n$ are a random sample from $N(\mu, \tau^{-1})$.
- First assume that τ is known and μ is unknown and that, as the prior for μ,

$$\mu \sim N(\mu_0, \tau_0^{-1}).$$

- Then,

$$\begin{aligned} p(\mu|D,\tau) &\propto L(D|\mu,\tau)q(\mu) \\ &\propto \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(Y_i-\mu)^2\right\}\exp\left\{-\frac{\tau_0}{2}(\mu-\mu_0)^2\right\} \\ &\propto \exp\left\{-\frac{n\tau}{2}(\mu^2-2\overline{Y}\mu)\right\}\exp\left\{-\frac{\tau_0}{2}(\mu^2-2\mu_0\mu)\right\} \\ &\propto \exp\left\{-\frac{n\tau+\tau_0}{2}\left(\mu-\frac{n\tau\overline{Y}+\tau_0\mu_0}{n\tau+\tau_0}\right)^2\right\}. \end{aligned}$$

Examples of deriving posterior distribution

- Hence,

$$\mu|D,\tau \sim N\left(\frac{n\tau\overline{Y}+\tau_0\mu_0}{n\tau+\tau_0}, (n\tau+\tau_0)^{-1}\right).$$

- Next, assume that μ is known and τ is unknown and that

$$\tau \sim \text{Gamma}(2^{-1}\delta_0, 2^{-1}\gamma_0),$$

  i.e., $q(\tau) \propto \tau^{\delta_0/2-1}\exp(-\gamma_0\tau/2)$.

- The posterior distribution of τ is

$$\begin{aligned} p(\tau|D,\mu) &\propto L(D|\mu,\tau)q(\tau) \\ &\propto \tau^{n/2}\exp\left\{-\frac{\tau\sum_{i=1}^n(Y_i-\mu)^2}{2}\right\}\tau^{\delta_0/2-1}\exp(-\gamma_0\tau/2) \\ &\propto \tau^{(n+\delta_0)/2-1}\exp\left\{-\frac{\tau\left(\sum_{i=1}^n(Y_i-\mu)^2+\gamma_0\right)}{2}\right\}. \end{aligned}$$

Examples of deriving posterior distribution

- So that

$$\tau|D,\mu \sim \text{Gamma}\left(\frac{n+\delta_0}{2}, \frac{\sum_{i=1}^{n}(Y_i-\mu)^2+\gamma_0}{2}\right).$$

- Finally, assume that both μ and τ are unknown. Specify the following prior $q(\mu,\tau) = q(\mu|\tau)q(\tau)$, where

$$\mu|\tau \sim N(\mu_0, \tau^{-1}\tau_0^{-1}),$$

and

$$\tau \sim \text{Gamma}(2^{-1}\delta_0, 2^{-1}\gamma_0).$$

It can be shown that

$$p(\pi,\tau|D) = p(\pi|\tau,D)p(\tau|D) = \text{Normal} \times \text{Gamma}.$$

The Bayesian Paradigm

- For the first example, a beta prior on θ leads to a beta posterior for θ. In the first scenario of the second example, a normal prior on μ yields a normal posterior for μ. In the second scenario of the second example, a gamma prior for τ yields a gamma posterior for τ.
- When the posterior distribution of a parameter is of the same family as the prior distribution, such prior distributions are called conjugate priors.
- Other conjugate priors include

Table 5.1: Examples of conjugate priors.

Likelihood	Conjugate prior
Poisson(θ)	Gamma(δ_0, γ_0)
Gamma(α, λ), α known	Gamma(δ_0, γ_0)
Beta($\alpha, 1$)	Gamma(δ_0, γ_0)

The Bayesian Paradigm

- In most cases, the posterior cannot be calculated analytically, and one generally has to resort to the Gibbs sampler.
- In §3.2 we described the Gibbs sampler as a tool to draw missing data from their conditional distribution given the observed data and the current iterate of θ in order to compute the conditional expectation at the E step.
- Recall that the idea in Gibbs sampling is to generate samples by sweeping through each variable (or block of variables) to sample from its conditional distribution with the remaining variables fixed to their current values.
- Draws from each one-dimensional conditional distribution are completed using the adaptive rejection sampling (Gilks and Wild, 1992).

The Bayesian Paradigm

- Specifically, suppose $\theta = (\theta_1, \cdots, \theta_p)^{\mathrm{T}}$. To draw θ from the posterior distribution $[\theta|D]$:
  1. Initialize $\theta_1^{(0)}, \cdots, \theta_p^{(0)}$;
  2. For $t = 1, 2, \cdot$, do
     - 2.1 $\theta_1^{(t)} \sim [\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \cdots, \theta_p^{(t-1)}, D]$
     - 2.2 $\theta_2^{(t)} \sim [\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \cdots, \theta_p^{(t-1)}, D]$
     - $\cdots$
     - 2.p $\theta_p^{(t)} \sim [\theta_p|\theta_1^{(t)}, \theta_2^{(t)}, \cdots, \theta_{p-1}^{(t)}, D]$
- Each step in 2 is completed by the adaptive rejection sampling, which requires only that we know the kernel of the conditional distribution.
- For the one-dimensional conditional distribution of θ_j, the kernel is $L(D|\theta)q(\theta)$, viewed as a function of θ_j.
- Because samples from the early iterations are not from the target posterior, it is common to discard these samples. The discarded iterations are often referred to as the "burn-in" period.

The Bayesian Paradigm

- For example, one can run the Gibbs sampler for 3000 iterations, discard the first 1000 iterations as burn-in, and use the remaining 2000 for posterior inference.
- These 2000 observations can be used to compute the posterior mean, mode, and the highest density regions (HDR).

The Bayesian Paradigm

- To minimize subjectivity in the choice of priors, it is desirable to use a non-informative prior.
- A prior is non-informative if the prior is "flat" relative to the likelihood function. Thus, it has minimal impact on the posterior distribution of θ.
- Other names for non-informative prior are reference prior, vague prior, or flat prior.
- In general, a non-informative prior is one which is dominated by the likelihood, that is, it is a prior which does not change very much over the region in which the likelihood is appreciable and does not assume large values outside that range.
- A commonly used non-informative prior is

$$q(\theta) \propto 1.$$

The Bayesian Paradigm

- This is an improper prior in the sense that

$$\int q(\theta) = \infty.$$

- Improper priors are sometimes useful because they may lead to proper posterior distributions, i.e.,

$$\int L(\theta|D)q(\theta)d\theta < \infty.$$

- However, there is another drawback of using the "uniform" prior of $q(\theta) \propto 1$, which is that it depends on the parameterization.
- Suppose ψ is another parameterization of θ. Under the "uniform" prior for θ, the prior for ψ is

$$q(\psi) \propto \left|\det \frac{d\theta}{d\psi}\right|,$$

which is different from the "uniform" prior for ψ.

The Bayesian Paradigm

- An alternative choice for non-informative priors that overcomes the non-uniqueness issue is to use

$$q(\theta) \propto \mathcal{I}(\theta)^{1/2},$$

  where $\mathcal{I}(\theta)$ is the information for θ based on one observation.

- In the case that θ is a vector

$$q(\theta) \propto \{\det \mathcal{I}(\theta)\}^{1/2}.$$

- For example, suppose that, given θ, $Y_1, \cdots, Y_n$ are a random sample from $N(\theta, 1)$. Then $\mathcal{I}(\theta) = 1$. So the information-based non-informative prior coincides with the "uniform" prior, and

$$p(\theta|D) \propto L(D|\theta) \propto \exp\left(-\frac{n(\theta - \overline{Y})^2}{2}\right).$$

The Bayesian Paradigm

- So the posterior is proper and

$$\theta|D \sim N(\overline{Y}, n^{-1}).$$

- As another example, suppose that, given θ, $Y_1, \cdots, Y_n$ are a random sample from Poisson(θ). Because $\mathcal{I}(\theta) = \theta^{-1}$, we take the prior to be

$$q(\theta) \propto \theta^{-1/2}.$$

Then,

$$p(\theta|D) \propto L(D|\theta)\theta^{-1/2} \propto \theta^{n\overline{Y}-1/2} \exp(-n\theta) = \theta^{n\overline{Y}+1/2-1} \exp(-n\theta).$$

So

$$\theta|D \sim \text{Gamma}(n\overline{Y} + 1/2, n).$$

The Bayesian Paradigm

- In the foregoing discussion, an experiment is conducted in which $D = (Y_1, \cdots, Y_n)$ are observed. The model is given by $L(D|\theta)$, with prior $q(\theta)$, and posterior

$$p(\theta|D) = \frac{L(D|\theta)q(\theta)}{\int L(D|\theta')q(\theta')d\theta'} \propto L(D|\theta)q(\theta).$$

- Now we wish to predict a future (independent) observation Y from this population, i.e., we wish to construct the Bayesian predictive distribution of Z.
- The posterior predictive distribution plays a crucial role in multiple imputation.
- The posterior predictive distribution of Y given D is

$$p(Y|D) = \int p_Y(Y|\theta)p(\theta|D)d\theta,$$

that is, the conditional density of Y given θ averaged over the posterior distribution of θ.

The Bayesian Paradigm

- In the example of $N(\theta, 1)$ and $q(\theta) \propto 1$, we had that

$$\theta|D \sim N(\overline{Y}, n^{-1}).$$

Since for another Y (independent of D) we have that $Y|\theta \sim N(\theta, 1)$, after some calculus, we find that

$$Y|D \sim N(\overline{Y}, 1 + n^{-1}).$$

- In general, the posterior distribution does not have an analytic form. So the following two steps are needed in order to draw Y from its predictive distribution. For $j = 1, 2, \cdots$,

1. Draw $\theta^{(j)}$ from $[\theta|D]$ (possibly using Gibbs sampler);
2. Draw $Y^{(j)}$ from $p_Y(Y|\theta^{(j)})$.

Then, $Y^{(1)}, Y^{(2)}, \cdots$ constitute a random sample from $[Y|D]$.

Contents

Bayesian Analysis

- With full data D, Bayesian inference is based on the posterior

$$[\theta|D]. \tag{5.1}$$

  if the posterior density $p(\theta|D)$ does not have a closed-form, the Gibbs sampler can be used to draw from (5.1).
- When missing data are present, denote $D = (D_{\text{obs}}, D_{\text{mis}})$. Then the Bayesian analysis is focused on $[\theta|D_{\text{obs}}]$, the kernel of which usually does not have a closed form.
- Assuming data are MAR,

$$\begin{aligned} p(\theta|D_{\text{obs}}) &\propto L(D_{\text{obs}}|\theta)q(\theta) \\ &= \left(\int L(D|\theta)dD_{\text{mis}}\right) q(\theta). \end{aligned}$$

Bayesian Analysis

- However, the problem can be solved using the idea of Gibbs sampling.
- Suppose we want to sample from $[U]$, the distribution of a random vector U, and suppose we know how to draw from the conditional distributions $[U|V]$ and $[V|U]$. Then, by iteratively sampling from the two conditionals, we eventually get a sample from the joint distribution $[U, V]$, denoted as $(U^{(1)}, V^{(1)}), (U^{(2)}, V^{(2)}), \cdots$.
- Notice that $U^{(j)}$, $j =, 1, 2, \cdots$, is a sample from the marginal distribution $[U]$.

Bayesian Analysis

- This entails the following iterative sampling for drawing from $[\theta|D_{\text{obs}}]$.
- First initialize $\theta = \theta^{(0)}$ (using, e.g., complete-case MLE). Then for the $(j+1)$th iteration:
  1. Draw $D_{\text{mis}}^{(j+1)}$ from $[D_{\text{mis}}|D_{\text{obs}}; \theta^{(j)}]$
  2. Draw $\theta^{(j+1)}$ from $[\theta|D_{\text{obs}}, D_{\text{mis}}^{(j+1)}]$

  where each step possibly entails one iteration of Gibbs sampling through individual components of the vector to be drawn.
- Observe that step 2 is the same as the sampling procedure in the full data scenario (5.1), which possibly involves a Gibbs sampler using the kernel $L(D^{(j+1)}|\theta)q(\theta)$, $D^{(j+1)} = (D_{\text{obs}}, D_{\text{mis}}^{(j+1)})$.
- So, Bayesian analysis with missing data only adds one more layer of sampling to the full-data framework, that is step 1, sampling from conditional distribution of missing data given the observed data and the previous draw of the parameter.

Bayesian Analysis

- Hence in the Bayesian framework, one can view the missing data as just one "nuisance" component of the parameter.
- However, the dimension of D_{mis} is proportional to the sample size, which may substantially increase the computational burden.
- Notice that step 1 is the same as drawing missing data from their conditional distribution in the E step of Monte Carlo EM and that the kernel of the conditional density is $L(D|\theta^{(j)})$ (viewed as a function of D_{mis}).

Bayesian regression analysis with MAR covariates

- We specialize this general approach to the problem of regression analysis with covariates that are missing at random.
- Let the conditional distribution of of response Y given covariates Z be $p(Y|Z;\beta)$, where β is the parameter of interest.
- In order to fully specify the full-data likelihood, we need to postulate a model for covariate distribution. Let $p(Z|\alpha)$ denote the marginal density of Z, where α is a set of nuisance parameters.
- Write $\theta = (\beta, \alpha)$. For prior construction for θ, it is convenient to decompose the prior as $q(\theta) = q(\beta|\alpha)q(\alpha)$, or even $q(\theta) = q(\beta)q(\alpha)$. The results typically will not vary much if the priors are chosen to be non-informative.

Bayesian regression analysis with MAR covariates

- The first option $q(\theta) = q(\beta|\alpha)q(\alpha)$ is appealing when using the "information principle" to construct the prior.
- Specifically, denote $\mathcal{I}(\beta|Z)$ as the information of β resulting from the conditional likelihood $p(Y|Z;\beta)$, that is,

$$\mathcal{I}(\beta|Z) = \int \left(\frac{\partial \log p(y|Z;\beta)}{\partial \beta} \right)^{\otimes 2} p(y|Z;\beta) dy.$$

Then, the marginal information of β is

$$\mathcal{I}(\beta|\alpha) = E[\mathcal{I}(\beta|Z)|\alpha] = \int \mathcal{I}(\beta|z) p(z|\alpha) dz. \tag{5.2}$$

Then, using the information principle, we may choose

$$q(\beta|\alpha) = \{\det \mathcal{I}(\beta|\alpha)\}^{1/2}.$$

Bayesian regression analysis with MAR covariates

- Sometimes calculating $\mathcal{I}(\beta|\alpha)$ through (5.2) is difficult because the integral may not be analytic.
- In such cases, one has to construct non-informative priors by other means.
- Write $D_Y = (Y_1, \cdots, Y_n)$ and $D_Z = (Z_1, \cdots, Z_n)$. Denote

$$L_{Y|Z}(D_Y|D_Z;\beta) = \prod_{i=1}^{n} p(Y_i|Z_i;\beta),$$

and

$$L_Z(D_Z|\alpha) = \prod_{i=1}^{n} p(Z_i|\alpha).$$

- Then,

$$L(D|\theta) = L_{Y|Z}(D_Y|D_Z;\beta)L_Z(D_Z|\alpha).$$

Bayesian regression analysis with MAR covariates

- Thus, the joint distribution of (D, θ) is

$$L(D|\theta)q(\theta) = L_{Y|Z}(D_Y|D_Z;\beta)L_Z(D_Z|\alpha)q(\beta|\alpha)q(\alpha).$$

- We describe the Gibbs sampling steps for D_{mis}, β, and α and the associated kernel functions for drawing from the conditional distributions. For simplicity, the iteration index is omitted. Write $Z_i = (Z_{i,\text{obs}}, Z_{i,\text{mis}})$ and $Z_{i,\text{mis}}$ is missing at random.
  1. For $i = 1, \cdots, n$, sample $Z_{i,\text{mis}}$ from $[Z_{i,\text{mis}}|Y_i, Z_{i,\text{obs}}; \theta]$, whose kernel is $p(Y_i|Z_{i,\text{obs}}, Z_{i,\text{mis}}; \beta)p(Z_{i,\text{obs}}, Z_{i,\text{mis}}|\alpha)$.
  2. Sample β from $[\beta|D;\alpha]$, whose kernel is $L_{Y|Z}(D_Y|D_Z;\beta)q(\beta|\alpha)$.
  3. Sample α from $[\alpha|D;\beta]$, whose kernel is $L_Z(D_Z|\alpha)q(\beta|\alpha)q(\alpha)$.

Example 1: Linear regression with continuous MAR covariates

- Let

$$Y = \beta^{\mathrm{T}} Z + \epsilon,$$

  where Z consists of 1 and a p-dimensional vector of continuous covariates and $\epsilon \sim N(0, \tau^{-1})$.
- Assume that the last p components of Z follow multivariate normal with mean μ and variance V^{-1}.
- Write $\theta = (\beta, \tau, \mu, V)$. We have seen how to compute the MLE of θ using EM algorithm (with explicit E and M steps). Now we conduct Bayesian inference.
- For prior construction, we first use the information principle to construct non-informative priors for (β, τ) given (μ, V).

Example 1: Linear regression with continuous MAR covariates

- By straightforward calculation,

$$\mathcal{I}(\beta, \tau|Z) = \begin{pmatrix} \tau Z^{\otimes 2} & 0 \\ 0 & 2^{-1}\tau^{-2} \end{pmatrix},$$

so that

$$\mathcal{I}(\beta, \tau|\alpha) = \begin{pmatrix} \tau E[Z^{\otimes 2}|\mu, V] & 0 \\ 0 & 2^{-1}\tau^{-2} \end{pmatrix}.$$

- Hence,

$$\det \mathcal{I}(\beta, \tau|\alpha) \propto \tau^{-2} \det\{\tau E[Z^{\otimes 2}|\mu, V]\} = \tau^{p-1} \det E[Z^{\otimes 2}|\mu, V],$$

where we recall that $E[Z^{\otimes 2}|\mu, V]$ is a $(p+1)$-dimensional square matrix.

Example 1: Linear regression with continuous MAR covariates

- It remains to calculate $\det E[Z^{\otimes 2}|\mu, V]$. Because we can write

$$Z \sim N\left\{\begin{pmatrix} 1 \\ \mu \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & V^{-1} \end{pmatrix}\right\}$$

  It is easy to see that

$$E[Z^{\otimes 2}|\mu, V] = \begin{pmatrix} 1 & \mu^{\mathrm{T}} \\ \mu & \mu^{\otimes 2} + V^{-1} \end{pmatrix}.$$

- To calculate its determinant, we use the general rule that

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A) \det(D - CA^{-1}B).$$

Example 1: Linear regression with continuous MAR covariates

- Therefore we have

$$\det E[Z^{\otimes 2}|\mu, V] = |V|^{-1},$$

  where $|A| = \det A$ for any matrix A.

- Thus,

$$q(\beta, \tau|\mu, V) \propto \{\det \mathcal{I}(\beta, \tau|\mu, V)\}^{\frac{1}{2}} \propto \tau^{\frac{p-1}{2}} |V|^{-\frac{1}{2}}.$$

  For the information of (μ, V), it can be shown, though not trivially, that

$$\mathcal{I}(\mu, V) \propto |V|^{-1}.$$

  and so $q(\mu, V) \propto |V|^{-1/2}$.

- So over all

$$q(\beta, \tau, \mu, V) \propto \tau^{\frac{p-1}{2}} |V|^{-1}.$$

Example 1: Linear regression with continuous MAR covariates

- Now, we consider Gibbs sampling with respect to D_{mis}, β, τ, μ, and V. Since the sampling of D_{mis} from its conditional distribution given D_{obs} and θ is the same as in the E step of an EM algorithm (in fact we have shown that the conditionals are multivariate normal), we focus on the conditional sampling of the parameter components.
- We have that, (letting $\widehat{\beta}$ be the least squares estimator for β)

$$\begin{aligned} L_{Y|Z}(D_Y|D_Z;\beta,\tau) &\propto \tau^{n/2}\exp\left(-\frac{\tau\sum_{i=1}^n(Y_i-\beta^{\mathrm{T}}Z_i)^2}{2}\right) \\ &=\propto \tau^{n/2}\exp\left(-\frac{\tau\sum_{i=1}^n(Y_i-\widehat{\beta}^{\mathrm{T}}Z_i)^2}{2}\right) \\ &\times\exp\left(-\frac{\tau(\beta-\widehat{\beta})^{\mathrm{T}}\sum_{i=1}^n Z^{\otimes 2}(\beta-\widehat{\beta})}{2}\right). \end{aligned}$$

Example 1: Linear regression with continuous MAR covariates

- Likewise,

$$
\begin{aligned}
& L_Z(D_Z|\mu, V) \\
\propto & |V|^{n/2} \exp\left(-\frac{\mathrm{tr}\left\{\sum_{i=1}^n (Z_i-\mu)^{\otimes 2} V\right\}}{2}\right) \\
= & |V|^{n/2} \exp\left(-\frac{\sum_{i=1}^n (Z_i-\overline{Z})^{\mathrm{T}} V (Z_i-\overline{Z}) + n(\mu-\overline{Z})^{\mathrm{T}} V (\mu-\overline{Z})}{2}\right).
\end{aligned}
$$

- It can be seen that

$$
\begin{aligned}
[\beta|D;\tau,\mu,V] &\propto \exp\left(-\frac{\tau(\beta-\widehat{\beta})^{\mathrm{T}} \sum_{i=1}^n Z^{\otimes 2}(\beta-\widehat{\beta})}{2}\right) \\
&\sim N\left\{\widehat{\beta}, \left(\tau \sum_{i=1}^n Z_i^{\otimes 2}\right)^{-1}\right\}.
\end{aligned}
$$

Example 1: Linear regression with continuous MAR covariates

- Also,

$$[\tau|D;\beta,\mu,V] \propto \tau^{\frac{n+p-1}{2}} \exp\left(-\frac{\tau\sum_{i=1}^{n}(Y_i-\beta^{\mathrm{T}}Z_i)^2}{2}\right)$$
$$\sim \text{Gamma}\left(\frac{n+p+1}{2}, \frac{\sum_{i=1}^{n}(Y_i-\beta^{\mathrm{T}}Z_i)^2}{2}\right).$$

- Further,

$$[\mu|D;\beta,\tau,V] \propto \exp\left(-\frac{n(\mu-\overline{Z})^{\mathrm{T}}V(\mu-\overline{Z})}{2}\right)$$
$$\sim N(\overline{Z}, n^{-1}V^{-1}).$$

Example 1: Linear regression with continuous MAR covariates

- Finally, to see what $[V|D, \beta, \tau, \mu]$ is, we need the following definition of distribution of positive definite matrices.

  Definition 5.1 (Wishart Distribution)

  A random positive definite $p \times p$ matrix V follows Wishart(δ, Σ) if its density is given by

  $$p(V|\delta, \Sigma) = 2^{-\delta p/2}|\Sigma|^{\delta/2}\Gamma_p(\delta/2)^{-1}|V|^{\frac{\delta-p-1}{2}} \exp\left(-tr(\Sigma V)/2\right),$$

  where Γ_p is the p-variate gamma function.

Example 1: Linear regression with continuous MAR covariates

- Then, we have that

$$[V|D,\beta,\tau,\mu] \propto |V|^{n/2-1}\exp\left(-\frac{\mathsf{tr}\left\{\sum_{i=1}^{n}(Z_i-\mu)^{\otimes 2}V\right\}}{2}\right)$$
$$\sim \mathsf{Wishart}\left(n+p-1,\sum_{i=1}^{n}(Z_i-\mu)^{\otimes 2}\right).$$

- All four conditional distributions of the components of the parameter have closed forms and can be directly sampled from.

Example 2: GLM with MAR covariates

- In GLM with missing covariates, the conditional distributions of parameters typicially do not have closed forms.
- For example, in GLM with canonical link function and $\phi = 1$, we have

$$L_{Y|Z}(D_Y|D_Z;\beta) \propto \prod_{i=1}^{n} \exp\left(\beta^{\mathrm{T}} Z_i Y_i - a(\beta^{\mathrm{T}} Z_i)\right)$$
$$= \exp\left(\beta^{\mathrm{T}} \sum_{i=1}^{n} Z_i Y_i - \sum_{i=1}^{n} a(\beta^{\mathrm{T}} Z_i)\right).$$

- It is hard to specify a prior $q(\beta)$ such that $L_{Y|Z}(D|\beta)q(\beta)$ has a recognizable kernel.
- Also, the information principle in constructing non-informative priors does not lead to straightforward calculation.

Example 2: GLM with MAR covariates

- This is because in this case

$$\mathcal{I}(\beta|Z) = \dot{\mu}(\beta^{\mathrm{T}} Z) Z^{\otimes 2},$$

  and for nonlinear link $\mu(\cdot)$, the expectation $E[\mathcal{I}(\beta|Z)|\alpha]$ typically does not have a closed form.
- In such cases, it is convenient to take $q(\beta|\alpha) \propto 1$ if this results in proper posteriors.
- Then, Gibbs sampler generally needs to sweep through each univariate component of β and α. In doing so, notice that

$$[\beta|D, \alpha] \propto L_{Y|Z}(D_Y|D_Z; \beta) \propto \exp\left(\beta^{\mathrm{T}} \sum_{i=1}^{n} Z_i Y_i - \sum_{i=1}^{n} a(\beta^{\mathrm{T}} Z_i)\right),$$

  and

$$[\alpha|D, \beta] \propto L_Z(D_Z|\alpha) q(\alpha).$$

Bayesian inference for non-ignorable missing data

- For missing data that are NMAR, a model for the missingness mechanism has to be specified, say, by

$$\pi(R|Y;\psi),$$

  where R is a variable indicating the missing pattern and ψ is a finite-dimensional parameter.
- The full data now are $D = \{(R_i, Y_i), i = 1, \cdots, n\}$, and the full data likelihood is

$$L(D|\theta) = \prod_{i=1}^{n} p_Y(Y_i|\theta)\pi(R_i|Y_i;\psi).$$

- Specify a joint prior for (θ, ψ) as $q(\theta, \psi)$.

Bayesian inference for non-ignorable missing data

- Then, for the Gibbs sampler, we iteratively draw from

$$[D_{\mathsf{mis}}|D_{\mathsf{obs}};\theta,\psi] \propto L(D|\theta),$$

$$[\theta|D;\psi] \propto q(\theta,\psi)\prod_{i=1}^{n} p_Y(Y_i|\theta),$$

and

$$[\psi|D;\psi,\theta] \propto q(\theta,\psi)\prod_{i=1}^{n} \pi(R_i|Y_i;\psi),$$

where the first step is identical to that in the E step of an MCEM with non-ignorable missing data.

Contents

Multiple Imputation

- The idea of multiple imputation is due to Rubin, who proposed it in several papers in the late 1970s. Since then, there has been an extensive body of work devoted to multiple imputation, including a review by Rubin (1996).
- Multiple imputation holds great practical appeal because its basic idea is very simple and intuitive.
- To conduct multiple imputation, the missing values are filled in, i.e., imputed, K times to create K "full" data sets.
- The full data analysis of interest is carried out on each of these K imputed data sets.
- The results of the K analyses are combined into a single analysis that takes into account the imputation.

Multiple Imputation

- Specifically, the procedure works as follows:
  1. Construct K full datasets by inserting in the missing values drawn from their conditional distribution given the observed values;
  2. Obtain an estimate $\widehat{\theta}^{(k)}$ for the kth imputed dataset (say, by MLE), $k = 1, ..., K$.
  3. The parameter estimate is $\widehat{\theta} = K^{-1} \sum_{k=1}^{K} \theta^{(k)}$.
  4. To compute the variance estimate, let $\widehat{V}^{(k)}$ denote the variance estimate from the kth imputed full dataset, obtained by, e.g., the inverse information matrix.
  5. Compute
     - Within imputation variation: $\overline{V} = K^{-1} \sum_{k=1}^{K} \widehat{V}^{(k)}$;
     - Between imputation variation: $\widehat{B} = K^{-1} \sum_{k=1}^{K} (\widehat{\theta}^{(k)} - \widehat{\theta})^{\otimes 2}$.
  6. The variance of $\widehat{\theta}$ is given by

$$\widehat{V}^{MI} = \overline{V} + (1 + K^{-1})\widehat{B}. \tag{5.3}$$

Multiple Imputation

- The underlying logic behind multiple imputation is as follows.
- Assume the data are MAR. If we knew the true value of θ, then we could generate full data D by first generating D_{obs} from $[D_{\text{obs}}|\theta]$ and then generating D_{obs} from $[D_{\text{mis}}|D_{\text{obs}};\theta]$.
- In the missing data context, the first step $[D_{\text{obs}}|\theta]$ is completed by nature. Multiple imputation is used to complete the second step $[D_{\text{mis}}|D_{\text{obs}};\theta]$. But because there is an objective uncertainty with the non-existent D_{mis}, imputation is done multiple times both to increase the precision of the final estimates and to quantify this uncertainty (by $\widehat{B}$).
- The fundamental question arises: since we do not know θ, how can we conduct the imputation based on $[D_{\text{mis}}|D_{\text{obs}};\theta]$?

Proper vs improper MI

- There are two types of MI schemes that can be used to implement the imputation step, improper and proper MI.
- With **improper** imputation, we start with an initial estimator $\widehat{\theta}^{init}$. This can be the MLE of CC analysis. Then, imputations are conducted using $[D_{\text{mis}}|D_{\text{obs}};\widehat{\theta}^{init}]$.
- There are two problems associated with with is approach.
- One is that $\widehat{\theta}^{init}$ may be inconsistent, i.e., does not converge to the true value in large sample sizes. Therefore the imputation distribution $[D_{\text{mis}}|D_{\text{obs}};\theta]$ is wrong, leading to biased inference.
- On the other hand, if $\widehat{\theta}^{init}$ is valid, the MI estimator will be valid. But in this case there would probably be less point in conducting the MI than using $\widehat{\theta}^{init}$ itself as the estimator.

Proper vs improper MI

- In fact, there are scenarios where a valid and efficient $\widehat{\theta}^{init}$ results in a less efficient MI estimator (Tsiati, 2006; Davidian, 2017).
- The other problem with the improper procedure is that (5.3) is not a correct variance estimator for the resulting MI estimate.
- This is because the imputation step is fixed at $\widehat{\theta}^{init}$, and so the uncertainty contained in $\widehat{\theta}^{init}$ is not properly reflected.
- A "sort-of" proper MI is, in the kth imputation step, first draw a parameter value $\theta^{*(k)}$ from

$$N(\widehat{\theta}^{init}, \widehat{\Sigma}^{init}),$$

where $\widehat{\Sigma}^{init}$ is some estimator for the variance of $\widehat{\theta}^{init}$, and then impute the missing data from $[D_{\text{mis}}|D_{\text{obs}}; \theta^{*(k)}]$.

Proper vs improper MI

- This "sort-of" proper MI addresses the second problem of improper MI by drawing a fresh θ for each imputation step to reflect its uncertainty.
- But the first problem remains because if $\widehat{\theta}^{init}$ is not valid, then the distribution of $N(\widehat{\theta}^{init}, \widehat{\Sigma}^{init})$ does not center around the true value of θ either.
- With **proper** imputation, the missing values are generated from a Bayesian posterior predictive distribution $[D_{\text{mis}}|D_{\text{obs}}] \propto \int p(D_{\text{mis}}|D_{\text{obs}};\theta)p(\theta|_{\text{obs}})d\theta$.
- The proper imputation is valid because the posterior distribution of θ, looked at from the frequentist vantage point, is always centered around its true value in large samples, no matter the choice of the prior.
- It also acknowledges the uncertain for the parameter in the imputation step because it is averaged across it posterior distribution.

Proper vs improper MI

- If there is a closed form of conditional distribution from which $[\theta|D_{\text{obs}}]$ can be drawn, then drawing from the posterior predictive distribution can be decomposed into two steps.
- First, we draw θ from $[\theta|D_{\text{obs}}]$ and then draw D_{mis} from $[D_{\text{mis}}|D_{\text{obs}};\theta]$.
- In general, to carry out proper MI, one needs the Gibbs sampler.
- Suppose the prior for θ is $q(\theta)$. Along the lines of §5.2, the following iterative sampling procedure can be used to draw from $[D_{\text{mis}}|D_{\text{obs}}]$.

Sampling from the posterior predictive distribution

- First initialize θ. Then, at the $(j+1)$th iteration,
  1. Draw $D_{\text{mis}}^{(j+1)}$ from $[D_{\text{mis}}|D_{\text{obs}};\theta^{(j)}] \propto L(D|\theta^{(j)})$;
  2. Draw $\theta^{(j+1)}$ from $[\theta|D_{\text{obs}}, D_{\text{mis}}^{(j+1)}] \propto L(D|\theta^{(j)})q(\theta)$.
- That is, the Gibbs sampling steps are exactly the same as the fully Bayesian analysis.
- The difference is: for Bayesian analysis, $\theta^{(M+1)}, \theta^{(M+2)}, \cdots$ are collected as a sample from the posterior distribution, where M is the number of iterations in the burn-in period. For MI, $D^{(M+1)}, D^{(M+2)}, \cdots$, where $D^{(j)} = (D_{\text{obs}}, D_{\text{mis}}^{(j)})$, $j = M+1, M+2, \cdots$, are collected as imputed datasets for frequentist analyses.

Imputation from a multivariate normal distribution

- Our development up to now has assumed that the model that is used to draw the imputations is the same as that to be used to analyze the data.
- That is, if the likelihood for full data analysis is $L(D|\theta)$, and the likelihood used to draw from $[\theta^*|D]$ and $[D_{\text{mis}}|D_{\text{obs}};\theta^*]$ is $L^{imp}(D|\theta^*)$, we have assumed that $L^{imp}(D|\theta) = L(D|\theta)$.
- Call $L(D|\theta)$ the analyst's model and $L^{imp}(D|\theta^*)$ the imputer's model.
- The term "congenial" coined by Meng (1994), has been used to describe the situation where analyst's model and the imputer's model are compatible, that is, $L^{imp}(D|\theta) = L(D|\theta)$.

Imputation from a multivariate normal distribution

- However, for general problems, it may not always be feasible or convenient to use congenial models, because it may be difficult to generate imputations in the context of a complex full data model $L(D|\theta)$.
- This has led to examination of performance of MI methods when the models are not congenial, that is, using a model such that $L^{imp}(D|\theta) \neq L(D|\theta)$.
- When all components of full data Y are continuous, a natural choice for $L^{imp}(D|\theta)$ is the multivariate normal.
- Interestingly, it has been shown via extensive numerical studies that, if the proportion of missing data is not great, this approach can lead to reasonable results, even if missing components of Y are not continuous but are categorical. Schafer (1997) offers extensive discussion.

Imputation from a multivariate normal distribution

- Accordingly, available software for multiple imputation, such as SAS proc MI and R packages such as NORM, AMELIA, or MICE, base imputation fully or in part on a multivariate normal model.
- We first demonstrate how imputation can be carried out assuming multivariate normality first in the case where missingness is monotone.
- We model the full data $Y = (Y_1, \cdots, Y_p)$ as follows: for $j = 1, \cdots, p$, assume

$$Y_j|(Y_1, \cdots, Y_{j-1}) = N\left(\beta_{j0} + \beta_{j1}Y_1 + \cdots + \beta_{j,j-1}Y_{j-1}, \tau_j^{-1}\right). \quad (5.4)$$

Imputation from a multivariate normal distribution

- Write $\beta_j = (\beta_{j0}, \cdots, \beta_{j,j-1})^{\mathrm{T}}$. Assume that missingness are monotone, i.e., if Y_j is missing then all $Y_{j'}$ are missing $\forall j' > j$,
- We can show that the posterior distribution for the β_j and τ_j given the observed data have closed forms. And because the conditional distributions of missing values given observed values and fixed parameters are multivariate normal, no Gibbs sampling is needed.
- Specifically, we can show that the observed data likelihood for β_j and τ_j involves only complete cases, i.e., subject that are "censored" later than the jth component of Y. This is not surprising because if one of the covariates in (5.4) is missing, then the response is also missing (by monotone missingness assumption), and then the contribution to the likelihood contains no regression parameter (see §1.2).

Imputation from a multivariate normal distribution

- If we choose the improper prior $q(\beta_1, \tau_1, \cdots, \beta_p, \tau_p) \propto \prod_{j=1}^{p} \tau_p^{-1}$, then, one can show

$$\tau_j | D_{\text{obs}} \sim \widehat{\tau}_j^2 \chi_{n_j - j} / n_j,$$

$$\beta_j | (D_{\text{obs}}; \tau_j) \sim N\left(\widehat{\beta}_j, \tau_j^{-1} (S_j^{\mathrm{T}} S_j)^{-1}\right),$$

  where n_j is the number of complete cases for the jth regression model, $\widehat{\beta}_j$ and $\widehat{\tau}_j$ are the CC (thus observed-data) MLEs, and S_j is the design matrix for the regression model for the regression model (5.4).
- Moreover, because of the factorization in the likelihood, the (β_j, τ_j) are independent given D_{obs}.
- So, to draw (β_j, τ_j) from the posterior, one first draws τ_j from $[\tau_j | D_{\text{obs}}]$ and then β_j from $[\beta_j | D_{\text{obs}}; \tau_j]$.

Imputation from a multivariate normal distribution

- Now, to complete the imputation, it remains to draw D_{mis} from $[D_{\text{mis}}|D_{\text{obs}};\theta^*]$, where $\theta^* = (\beta_1, \tau_1, \cdots, \beta_p, \tau_p)$ are freshly drawn from their posterior.
- Because of the special structure, this can be done progressively in $j = 1, \cdots, p$.
- For each subject, assume the missing values in $Y_1, \cdots, Y_{j-1}$ have been imputed. Then, because of the imputation model (5.4), Y_j can be imputed by

$$Y_j = \beta_{j0} + \beta_{j1}Y_1 + \cdots + \beta_{j,j-1}Y_{j-1} + \tau_j^{-1/2}\epsilon_j,$$

where ϵ_j is simulated from $N(0,1)$.

Imputation from a multivariate normal distribution

- In the case with nonmonotone missingness, Gibbs sampling has to be employed.
- However, if we still assume a multivariate normal imputation model $N(\mu^*, \Sigma^*)$, then typically the Gibbs sampler consists of iterative draws of D_{mis}, μ^*, and Σ^*, where each (conditional) distribution has an analytic form (along the lines of Example 1 in §5.2). So, one does not need to sweep through each univariate component of the parameters and missing values.
- MI based on multivariate imputation models for monotone and nonmonotone missing data are implemented by the SAS PROC MI procedure.
- A hands-on introduction to the SAS procedure is given in Yuan (2011).

Multivariate imputation by chained equations (MICE)

- As noted above, when the full data comprise both continuous and categorial variables, one strategy is just to proceed as if all were continuous and approximately normal and generate imputations from a multivariate normal distribution.
- An alternative approach to adopting an imputation model for the entire joint distribution of the full data is to adopt a fully conditional specification (FCS), also known as multivariate imputation by chained equations (MICE) (van Buuren, 2007; van Buuren and Groothuis-Oudshoorn, 2011).
- This multiple imputation approach using chained equations has gained considerable recent popularity.

Multivariate imputation by chained equations (MICE)

- We describe the basic idea of the MICE algorithm as follows.
- Assume the full data for one individual is $Y = (Y_1, \cdots, Y_p)$. First, initialize by filling in all missing values for each individual (say, by random sampling from the observed values).
- The first variable with missing values for some individuals, say Y_1, is regressed on all other variables $Y_2, \cdots, Y_p$, restricting to individuals for whom Y_1 is observed. Then, Y_1 for individuals for whom it is is missing are imputed via simulated draws from the corresponding "posterior predictive distribution" of Y_1. (The regression model of course takes into account the nature of Y_1, e.g., linear regression, logistic, or Poisson regression if Y_1 is continuous, binary, or count data, respectively).

Multivariate imputation by chained equations (MICE)

- Do the same thing for Y_j, i.e., regress on all other variables based on the observed Y_j, and then fill in the missing values from the "posterior predictive distribution".
- This scheme continues sequentially for all other variables with missing values for some individuals. This completes one cycle of the algorithm.
- This process for all variables is repeated for several cycles (e.g., 10 to 20) to stabilize the result. This results in a **single** imputed data set. The entire procedure is carried out K times to yield K imputed data sets.
- The term chained equations thus refers to the successive regression equations (models) used in the sequence for each cycle. They are "chained" because each Y_j is modeled on all the other components of Y.

Multivariate imputation by chained equations (MICE)

- Now we look at what precisely is done in the imputation for the missing Y_j.
- If Y_j is continuous, then the regression model can be chosen to be a linear regression, and the imputation is similar to what's been described in the imputation step for the monotone missing case with multivariate normal model.
- If Y_j is categorical, then typically a GLM is used to model it, say, logistic regression. Denote the MLE of the regression parameter as $\widehat{\beta}_j$. Then, drawing from the "posterior predictive distribution" is similar to what we have mentioned as the "sort-of" proper imputation.
- Specifically, we first draw β_j from

$$N(\widehat{\beta}_j, \widehat{\Sigma}_j),$$

where $\widehat{\Sigma}_j$ is an estimate for the variance of $\widehat{\beta}_j$.

Multivariate imputation by chained equations (MICE)

- Then, Y_j is drawn from

$$\text{Bin}\left(1, \frac{\exp(\beta_j^{\mathrm{T}} Z_j)}{1+\exp(\beta_j^{\mathrm{T}} Z_j)}\right),$$

  where $Z_j = (Y_1, \cdots, Y_{j-1}, Y_{j+1}, Y_p)^{\mathrm{T}}$.
- The MICE algorithm is implemented in, for example, SAS PROC MI using the FCS option and the R package MICE.
- Some caveats: The MICE/FCS approach is predicated on specification of full conditional models; that is, in our example for each j, models for Y_j as a function of all other variables.

Multivariate imputation by chained equations (MICE)

- Clearly, it is virtually impossible in general, for variables of mixed types, to specify an entire set of such models such that they are all compatible.
- By this we mean that they all are consistent with a single joint distribution specification for all variables.
- Accordingly, there is no reason to expect, as is the case with Gibbs sampling methods based on a compatible set of conditional models derived from a joint distribution, that a sequence of imputed values generated by the algorithm need converge to anything.
- We thus emphasize again that the algorithm is *ad hoc*, although it seems to work well in practice.

MI for non-ignorable missing data

- When data are NMAR, we need to postulate a selection model, say, $\pi(R|Y;\psi)$. Then, the imputation step needs to be based on the (imputation) full-data likelihood

$$L^{imp}(D|\theta^*)\prod_{i=1}^{n}\pi(R_i|Y_i;\psi),$$

  and one has to specify a joint prior for (θ^*,ψ) as in the case of fully Bayesian analysis.
- Obviously, whether the data are MAR only affects the imputation step and does not affect the analysis step, i.e., it is of no concern to the one who analyzes the imputed datasets.
- This is one advantage of MI over the competing methods.

Concluding remarks on MI

- Multiple imputation is a popular approach to handling missing data, owing to the fact that "filling in" missing values has great practical appeal.
- The original literature on multiple imputation suggested that $K = 3$ to 5 imputed data sets are all that may be needed to obtain good results.
- Studies of this issue by simulation have shown that a much larger number M of imputed data sets should be used to obtain accurate estimators of uncertainty.
- Recommendations of $K = 30$ to 50 have been made on this basis more recently, and an *ad hoc* rule of thumb that has been proposed is to take K to be roughly equal to the percentage of missing information.

References

- Davidian, m. (2017) Lecture notes: Statistical methods for analysis with missing data. http://www4.stat.ncsu.edu/~davidian/st790/notes.html.
- Gilks, W. R. & Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. Applied Statistics, 337-348.
- Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. Statistical Science, 538-558.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American statistical Association, 91, 473-489.
- Schafer, J. L. (1997). Analysis of incomplete multivariate data. CRC press.
- Tsiatis, A. (2006). Semiparametric theory and missing data. New York: Springer.
- van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Statistical methods in medical research, 16, 219-242.
- van Buuren, S. & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. Journal of statistical software, 45.
- Yuan, Y. (2011). Multiple imputation using SAS software. Journal of Statistical Software, 45, 1-25.