2000 Proceedings of the Autumn Conference, Korean Statistical Society

A nonnormal Bayesian imputation

Minwoong Shin, Jinhee Lee[1], Juyoung Lee[2], Sangeun Lee[3]

Abstract

When the standard inference is to be used with complete data and nonresponse is ignorable, then multiple imputations should be created as repetitions under a Bayesian normal model. Many Bayesian models besides the normal, however, approximately yield the standard inference with complete data and thus many such models can be used to create proper imputations. We consider the Bayesian bootstrap(BB) application.

1. Introduction

In sample survey, for the imputation methods, in general, there are 8 principal ones : hot-dect, substitution, cold-dect, regression, stochastic regression, composite and multiple imputation. Recently, in literature, multiple imputation method which was introduced by Rubin(1977) has been more payed attention to be use of among imputation methods. Multiple imputation refers to the statistical technique that substitute each missing or deficient values with two or more acceptable values representing distribution of possibilities and can be used to handle missing data in survey nonresponse context. Furthermore most of methods arises from the Bayesian perspective which are not only provides a simple and general theorical rational but also provides prescriptions for how to create multiple imputation and analysis the resultant data in specific areas.

In particular, if the multiple imputation are repetitively drawn to simulate a Bayesian Posterior distribution of the missing values under a model, the approximately combining analysis of each data set completed by imputation yields an approximately valid Bayesian inference under a model.

In this study. it will be proposed the multiple imputation with non-normal Bayesian repeated imputation procedure.

1) Department of Statistics, Hankuk University of Foreign Studies, Kyonggi-do, Korea, 449-791.
2) Division of Epidemiology, Department of Infectious Disease, National Institute of Health, #5 Nokbun-dong, Eunpyung-gu, Seoul, Korea, 112-701
3) Department of Applied Information Statistics, Kyonggi University, Kyonggi-do, Korea, 442-760.

2. The distribution of finite population mean

We will let the number of units in the finite population be denoted by N. We will let X refer to fully observed covariates, such as stratum indicators or size of unit measurements, recorded for all N units in the population :

$$X=(X_1,\cdots,X_N)' \tag{2.1}$$

where X_i will be a row vector, if there exists more than are fully observed covariate.

For example, letting the units be people, where $X_i=(X_{i1},X_{i2},X_{i3})$, X_{i1} could indicate the gender of the ith unit.

We will let Y refer to variables whose values are not known for all units in the populations :

$$Y=(Y_1,Y_2,\cdots,Y_N)' \tag{2.2}$$

The variable

$$I=(I_1,I_2,\cdots,I_N)' \tag{2.3}$$

will be the indicator for inclusion/exclusion from the survey.

The variable

$$R=(R_1,R_2,\cdots,R_N)' \tag{2.4}$$

will be the indicator for respondent or nonrespondent. In the case of joint one outcome variable, R_i is binary with $R_i=1$ indicating that ith unit will respond and $R_i=0$ indicating that the ith unit will not respond.

Suppose that for $i=1,2,\cdots,N$ the distribution of Y_i given $\theta=(\mu,\sigma^2)$ is i.i.d $N(\mu,\sigma^2)$ and the proportional to σ^{-2}.

Then the distribution of $\overline{Y}=\sum_{i=1}^{N} Y_i/N$ given $Y_1,Y_2,\cdots,Y_N$ is a t on $n-1$ degrees of freedom with location $\overline{y}$ and scale $s(n^{-1}-N^{-1})^{1/2}$ where $\overline{y}=\sum y_i/n$ and $s^2=\sum(Y_i-\overline{y})^2/(n-1)$.

Lemma 2.1

Suppose that given (μ,σ^2), Y_i, $i=1,2,\cdots,N$ are i.i.d. $N(\mu,\sigma^2)$ and that the marginal(prior) distribution of (μ,σ^2) has density proportional to σ^{-2}, that is, suppose

$$P_r(Y_1,Y_2,\cdots,Y_N\mid\mu,\sigma^2)=\Pi_{i=1}^{N}(2\pi\sigma^2)^{-1/2}\exp[-(Y_i-\mu)^2/2\sigma^2] \tag{2.5}$$

and

$$P_r(\mu,\sigma^2)\propto\sigma^{-2}. \tag{2.6}$$

Then (i) the distribution of μ given $(\sigma^2,Y_1,Y_2,\cdots,Y_n)$ is $N(\overline{y},\sigma^2/n)$, where

$\bar{y}=\sum Y_i/n$, that is $P_r[\mu \mid \sigma^2, Y_1, Y_2, \cdots, Y_n]=(2\pi\sigma^2/n)^{-1/2}\exp[-(\mu-\bar{y})^2/(2\sigma^2/n)]$

(ii) the distribution of σ^2 given $Y_1, Y_2, \cdots, Y_n$ is $(n-1)s^2\chi_{n-1}^{-2}$ where $s^2=\sum(Y_i-\bar{y})^2/(n-1)$ and χ_{n-1}^{-2} is the inverted chi-square distribution with $n-1$ degree of freedom. Again suppose that Y_i is scalar and drop the assumption of normality. That is, suppose that given θ, the $Y_1, Y_2, \cdots, Y_n$ are i.i.d. where θ has prior distribution $p(\theta)$.

Let $E(Y_i \mid \theta)=\mu$ and $V(Y_i \mid \theta)=\sigma^2$, both finite function of θ.

Then

$$E(\bar{Y} \mid Y_1 Y_2, \cdots, Y_n)=(n/N)\bar{y}+(1-n/N)\hat{\mu} \tag{2.7}$$

where $\hat{\mu}=E(\mu \mid Y_1, Y_2, \cdots, Y_n)$

and

$$V(\bar{Y} \mid Y_1, Y_2, \cdots, Y_n)=(1-n/N)[\hat{\sigma}^2/N+(1-n/N)]\cdot V(\mu \mid Y_1, Y_2, \cdots, Y_n) \tag{2.8}$$

where $\hat{\sigma}^2=E[\sigma^2 \mid Y_1, Y_2, \cdots, Y_n]$.

2.1 The Bayesian Bootstrap

Let $d=(d_1, d_2, \cdots, d_k)$ be the vector of all possible district values of Y_i and let $\theta=(\theta_1, \theta_2, \cdots, \theta_k)$ be the associated vector of possibilities, $\sum\theta_i=1$, where we suppose that $Y_1, Y_2, \cdots, Y_n$ given θ are i.i.d.

$$P_r(Y_i=d_k \mid \theta)=\theta_k \tag{2.9}$$

Suppose further the prior distribution on θ is improper;

$$P_r(\theta)=\begin{cases} \Pi_{k=1}^{K}\theta_k^{-1}, & \text{if } \sum\theta_k=1 \\ 0, & \text{otherwise} \end{cases} \tag{2.10}$$

Let n_k= the number of $Y_i, i=1,2,\cdots,n$ equal to d_k, $\sum_{k=1}^{K} n_k=n$. The conditional distribution of θ given $Y_1, Y_2, \cdots, Y_n$, $P_r(\theta \mid Y_1, Y_2, \cdots, Y_n)$, is then the $(K-1)-$variate Dirichlet distribution proportional.

$$P_r(\theta)=\begin{cases} \Pi_{k=1}^{K}\theta_k^{-1}, & \text{if } \sum\theta_k=1 \text{ and } \sum n_k=n \\ 0, & \text{otherwise} \end{cases} \tag{2.11}$$

under this specification, values d_k that are not observed have zero probability given $(Y_1, Y_2, \cdots, Y_n)$.

The mean of Y_i given θ is

$$\mu = \sum_{k=1}^{K} d_k \theta_k \tag{2.12}$$

and the variance of Y_i given θ is

$$\sigma^2 \sum_{k=1}^{K} d_k^2 \theta_k - \mu^2 \tag{2.13}$$

The conditional mean of $\overline{Y}$ given $(Y_1, Y_2, \cdots, Y_n)$ is $\overline{y}$. The conditional variance of $\overline{Y}$ given $(Y_1, Y_2, \cdots, Y_n)$ is

$$\begin{aligned} V(\overline{Y} \mid Y_1, Y_2, \cdots, Y_n) &= (1-n/N)[\hat{\sigma}^2/N + (1-n/N)V(\mu \mid Y_1, Y_2, \cdots, Y_n)] \\ &= (1-n/N)[(n-1)s^2/nN + V(\mu \mid Y_1, Y_2, \cdots, Y_n) \\ &\quad \cdot (N-n-1)/N] \\ &= s^2(n^{-1} - N^{-1})(n-1)/(n+1) \end{aligned} \tag{2.15}$$

3. The Normal Model

The objective is to assess the posterior probability that the population mean $\overline{Y} = \sum_{i}^{N} Y_i/N$ is positive from a SRS of values of Y_i.

Thus $Q = \overline{Y}$ and the regression of interest for Q is $C = (0, \infty)$.

$$Q \mid Y_{\text{inc}} \sim t[(n-1), \overline{y}, s^2(n^{-1} - N^{-1})] \tag{3.1}$$

Suppose that due to unconfounded nonresponse, only n_1 of the values in Y_{inc} are observed. Let $\overline{y}_1$ and s_1^2 be the sample mean variance of the sampling and response mechanism are unconfounded and thus ignorable, it is easy to show that

$$\text{Prob}\{\overline{Y} > 0 \mid X, Y_{\text{inc}}, R_{\text{inc}}, I\} = \text{Prob}\{\overline{Y} > 0 \mid X, Y_{\text{inc}}\} \tag{3.2}$$

and

$$\text{Prob}\{\overline{Y} > 0 \mid X, Y_{\text{obs}}, R_{\text{inc}}, I\} = \text{Prob}\{\overline{Y} > 0 \mid X, Y_{\text{obs}}\} \tag{3.3}$$

Lemma 2.1 implies that the m draws from the posterior distribution of Y_{mis} (i.e. the distribution of the Y_{mis} given Y_{obs}) can be made as follows.

For $l = 1, 2, \cdots, m$, pass through the following three steps using independent draws for all random variables at each pass:

1. Draw a $\chi^2_{n_1 - 1}$ random variable, say x, and let

$$\sigma_*^2 = s_1^2(n_1 - 1)/x \tag{3.4}$$

2. Draw a

$N(0, 1)$ random variable, say z_0, and let

$$\mu_*^2 = \bar{y}_1 + \sigma_* z_0 / n_1 \tag{3.5}$$

3. Draw $n - n_1$ independent $N(0,1)$ random variable, say z_i, $i \in$ mis and impute the missing components of Y_{inc} as

$$Y_{i*} = \mu_* + \sigma_* z_i, \quad i \in \text{mis}. \tag{3.6}$$

Each of the m draws of Y_{mis} from its posterior distribution creates a computed value of Y_{inc} from $\text{Prob}\{\bar{Y} > 0 \mid X, Y_{\text{inc}}\}$ can be calculate as a t integral.

4. A nonnormal Bayesian imputation procedure

Consider the fully normal repeated-imputation method. It is easy to see that $\bar{Q}_\infty = \bar{y}_1$. For large samples, the average of the complete-data variances equals s_1^2.
In large samples

$$B_\infty = s_1^2 (n_1^{-1} - n^{-1}) \tag{4.1}$$

which is unbiased for B with low-order variability.

Fully normal Bayesian repeated imputation is proper for $\{\bar{y}_1, s^2(n_1^{-1} - N^{-1})\}$ in large samples with ignorable nonresponse for any popular of Y values. When the standard inference is to be used with complete data and nonresponse is ignorable then multiple imputations should be created as repetitions under a bayesian normal model.

Many models beside the standard inference with he complete data and thus many such models can be used to create proper imputations for $\{\bar{y}, s^2(n^{-1} - N^{-1})\}$ with ignorable nonresponse.

To illustrate this, consider the Bayesian bootstrap(BB) specification. A simple way to generate repeated imputation of Y_{mis} is to repeat the following two steps m independent times. We let $Y_{\text{obs}} = (Y_1, Y_2, \cdots, Y_{n_1})$

step1. draw $n_1 - 1$ uniform random numbers between 0 and 1 let their ordered values be $a_1, a_2, \cdots, a_{n_1 - 1}$; also let $a_0 = 0$ and $a_{n_1 - 1} = 1$.

step2. Draw each of the n_0 missing values in Y_{mis} by drawing from $Y_1, Y_2, \cdots, Y_{n_1}$ with probabilities $(a_1 - a_0), (a_2 - a_1), \cdots, (1 - a_{n_1 - 1})$; that is, independently n_0 times, draw a uniform random number u and impute Y_i if $a_{i-1} < u < a_i$.

Reference

Cochran(1963) W. G. (1963). Sampling Techniques. New York, Wiley.

Little R J. A and Rubin D. B(1987). Statistical analysis with missing data. John Wiley & Sons.

Lohr S. L(1999). Sampling : Design and analysis ; Duxbury Press.

Rubin D. B(1987). Multiple imputation for nonresponse in surveys. John Wiley & Sons.