A Bayesian Analysis of a Proportion under Nonignorable Nonrsponse

Balgobin Nandram
Mathematical Sciences, Worcester Polytechnic Institute,
100 Institute Road, Worcester, MA 01609-2280, USA
(balnan@wpi.edu)

Jai Won Choi
National Center for Health Statistics
Room 915, 6525 Belcrest Road, Hyattsville, MD 20782, USA
(jwc7@cdc.gov)

Methodology for Nonignorable Nonresponse

We use a hierarchical Bayesian selection model to accommodate the nonresponse mechanism. Our main result is that for some of the states the nonresponse mechanism can be considered nonignorable, and that 95% credible intervals of the probability for a household doctor visit shed important light on the nonresponse of the National Health Interview Survey (NHIS) data. We describe the expansion model and to show how to fit it using a full Bayesian method. Below we describe our model.

Let ℓ be the number of states, we assume that a sample of n_i households is taken from the i^{th} area, $i = 1, \ldots, \ell$. Let the binary characteristic be $y_{ij} = l$ if at least one member of household j in area i visited doctor and $= 0$ otherwise. The response variable $r_{ij} = 1$ if household j in area i is a respondent. We use a probabilistic structure to model y_{ij} and r_{ij}. Let $r_i = \sum_{j=1}^{n_i} r_{ij}$ be the number of households with respondents and $y_i = \sum_{j=1}^{n_i} y_{ij}$ the number of households with at least one doctor visit, and $n_i - r_i$ is the number of nonrespondents.

The expansion model for nonignorable nonresponse is $y_{ij} \mid p_i$ is $Bernoulli(p_i)$, $r_{ij} \mid \pi_i, \gamma_i, y_{ij} = 1$ is $Bernoulli(\gamma_i \pi_i)$, and $r_{ij} \mid \pi_i, \gamma_i, y_{ij} = 0$ is $Bernoulli(\pi_i)$. The γ_i are the ratios of the odds of success (doctor visit) among respondents to the odds of success (doctor visit) among all individuals in the i^{th} area. The γ_i shows the extent of nonignorability of the nonrespondents and incorporate the uncertainty about ignorability into the model. If $\gamma_i = 1$, the model becomes ignorable and there is no difference between respondents and nonrespondents.

The parameters of interest are p_i, γ_i and δ_i where δ_i is the probability of responding in area i and is given by $\delta_i = \pi_i\{\gamma_i p_i + (1-p_i)\}$, $p_i \mid \mu_1, \tau_1$ is $Beta(\mu_1, \tau_1, (1-\mu_1)\tau_1)$. We wish to center the γ_i at unity (i.e., center on an ignorable model). It is possible to do so by assuming that the γ_i have a common mean of unity. Thus, one can assume that $\gamma_i \mid \nu$ is approximately $\Gamma(\nu, \nu)$. Thus, we assume that the parameters (π_i, γ_i) are jointly independent with $\pi_i \mid \mu_2, \tau_2$ is $Beta(\mu_2\tau_2, (1-\mu_2)\tau_2)$ and $\gamma_i \mid \nu, \pi_i$ is $\Gamma(\nu, \nu)$ $0 < \gamma_i < 1/\pi_i$ and $0 < \pi_i < 1$.

For a full Bayesian analysis, we take some prior densities, μ_r is $Beta(1,1)$ $r = 1, 2$, and $p(\nu) = 1/(\nu+1)^2$, $\nu \geq 0$ and $p(\tau_r) = 1/(\tau_r+1)^2$, $\tau_r \geq 0$, $r = 1,\ 2$. Since the number of visits among the nonrespondents is unknown, we denote it by the latent variable $z_i.$, and hence, the number of households with no visits among them is $n_i - r_i - z_i$.

Using Bayes' theorem, transforming $\phi_i = \gamma_i \pi_i$, the joint posterior density of all the

parameters $(\mathbf{z}, \mathbf{p}, \boldsymbol{\pi}, \boldsymbol{\phi}, \mu_1, \tau_1, \mu_2, \tau_2, \nu)$ for given data $(\mathbf{y}, \mathbf{r})$ is posterior density

$$f(\mathbf{p}, \boldsymbol{\phi}, \boldsymbol{\pi}, \mathbf{z}, \mu_1, \tau_1, \mu_2, \tau_2, \nu \mid \mathbf{y}, \mathbf{r})$$
$$\propto p(\nu)p(\tau_1)p(\tau_2) \prod_{i=1}^{\ell} \left\{ \begin{pmatrix} n_i - r_i \\ z_i \end{pmatrix} \frac{p_i^{y_i+z_i+\mu_1\tau_1-1} (1-p_i)^{n_i-y_i-z_i+(1-\mu_1)\tau_1-1}}{B(\mu_1\tau_1,\ (1-\mu_1)\tau_1)} \phi_i^{y_i+\nu-1}(1-\phi_i)^{z_i} \right\}$$

$$\times \prod_{i=1}^{\ell} \left\{ \frac{\pi_i^{r_i-y_i+\mu_2\tau_2-1} (1-\pi_i)^{n_i-r_i-z_i+(1-\mu_2)\tau_2-1}}{B(\mu_2\tau_2,\ (1-\mu_2)\tau_2)} \left\{\pi_i^{-1} exp(-\phi_i/\pi_i)\right\}^{\nu} \nu\ I_i^{-1}(\mu_2, \tau_2, \nu) \right\}. \quad (1)$$

Inference about p_i, δ_i and γ_i can be obtained by using (**??**), but because this posterior density is complex, we use Markov chain Monte Carlo (MCMC) methods.

Analysis of NHIS Data

One of the variables we use in the NHIS is the number of doctor visits by the members of an entire household in the past two weeks. We use the binary variable, *doctor visit*, which is 0 if the number of doctor visits by all members of a household is 0, and 1 otherwise. For 50 states and the District of Columbia, we observed n_i, r_i and y_i which are the numbers of sampled households, responding households and households with doctor visits respectively. The observed proportion $\hat{\delta}_i = r_i/n_i$ of responding households among sample households ranges from 0.874 in the District of Columbia to 0.993 in Idaho. The ten states are the ones with smallest response rates (at least 7.2% nonrespondents) which we use for studying our method later. These states are Colorado, Delaware, District of Columbia, Florida, Louisiana, Maryland, Nevada, New York, South Carolina and West Virginia.

We apply our methodology to the NHIS data. we observed 95% credible intervals for the p_i, δ_i and γ_i. Some of the intervals do not contain the observed values of the p_i. For example, Colorado, Florida and New York do not contain the observed proportions. This implies that the ratio method may provide unreasonable estimates for the true proportions for these states. As for the δ_i there are some variations among the states, where for the 95% credible intervals the upper ends are reasonably close, but the lower ends differ. But in general the response rates are similar. It is clear that many of the intervals for the γ_i do not contain 1, and so for these states the nonresponse mechanism should be considered nonignorable, and therefore the ratio estimator should not be used.

RESUME

There is a substantial number of nonrespondents among the sampled households in the National Health Interview Survey (NHIS). The main issue we address here is that the nonresponse mechanism should not be ignored because respondents and nonrespondents differ. The purpose of this work is to estimate the proportion of households with at least one doctor visit, and to investigate what adjustment to be made for nonignorable nonresponse. We use a hierarchical Bayesian selection model to accommodate this nonresponse mechanism. Our main result is that for some of the states the nonresponse mechanism can be considered nonignorable, and that 95% credible intervals of the probability for a household doctor visit and the probability that a household responds shed important light on the NHIS data.