Journal of Official Statistics, Vol. 23, No. 4, 2007, pp. 433–452

Non-Bayesian Multiple Imputation

Jan F. Bjørnstad[1]

Multiple imputation is a method specifically designed for variance estimation in the presence of missing data. Rubin's combination formula requires that the imputation method is "proper," which essentially means that the imputations are random draws from a posterior distribution in a Bayesian framework. In national statistical institutes (NSI's) like Statistics Norway, the methods used for imputing for nonresponse are typically non-Bayesian, e.g., some kind of stratified hot-deck. Hence, Rubin's method of multiple imputation is not valid and cannot be applied in NSI's. This article deals with the problem of deriving an alternative combination formula that can be applied for imputation methods typically used in NSI's and suggests an approach for studying this problem. Alternative combination formulas are derived for certain response mechanisms and hot-deck type imputation methods.

Key words: Variance estimation; survey sampling; stratified sampling; logistic regression; nonresponse; hot-deck imputation.

1. Introduction

Multiple imputation is a method specifically designed for variance estimation in the presence of missing data, developed by Rubin (1987). Two more recent references with further discussions and studies are Rubin (1996) and Schafer (1997). The basic idea is to create m imputed values for each missing value and combine the m completed data sets by Rubin's combination formula for variance estimation. For the estimator to be valid, the imputations must display an appropriate level of variability. In Rubin's term, the imputation method is required to be "proper." In national statistical institutes (NSI's) the methods used for imputing for nonresponse very seldom if ever satisfy the requirement of being "proper." However, the idea of creating multiple imputations to measure the imputation uncertainty and use it for variance estimation and for computing confidence intervals is still of interest. The problem is then that Rubin's combination formula is no longer valid with the usual nonproper imputations used by NSI's. The reason is that the variability in nonproper imputations is too small and the between-imputation component must be given a larger weight in the variance estimate. The problem is then to determine what this weight should be to give valid statistical inference, and also for what kind of nonresponse mechanisms and estimation problems it is possible to determine a simple

[1] Statistics Norway, Division for Statistical Methods and Standards, P.O. Box 8131 Dep., N-0033 Oslo, Norway. Email: jab@ssb.no

Acknowledgment: The problem of deriving a non-Bayesian multiple imputation method was studied in a Master's thesis in 1999 by Tonje Braaten with the author as her adviser. The present research began within the DACSEIS research project, and started originally when the author was contacted by Tonje Braaten regarding this issue in her doctoral studies in epidemiology.

combination formula not dependent on unknown parameters. This article suggests an approach for studying this problem.

In Section 2 an approach for determining the combination of the imputed completed data sets is suggested. Section 3 has three applications with random nonresponse: (i) estimating a population average from simple random samples using hot-deck imputation, (ii) estimating the regression coefficient in the ratio model using residual regression imputation and (iii) estimating the regression coefficient in simple linear regression with residual regression imputation. Section 4 deals with the general problem of multiple imputation for stratified samples. In Section 5 we apply the theory in Section 4 to stratified samples with random nonresponse within strata, covering (i) estimation of population average using stratified hot-deck imputation and (ii) estimation of log (odds ratios) in logistic regression with missingness both for the dependent variable and the explanatory variable. Section 6 takes up the problem of using the same combination rule for all estimation problems with a given imputation method and data and response model. A general result for hot-deck imputation and linear estimates is presented.

2. An Approach for Determining an Alternative Combination Formula for Variance Estimation in Multiple Imputation

Let $s = (1, \ldots, n)$ denote the full sample, with $\mathbf{y} = (y_1, \ldots, y_n)$ denoting the full sample data, values of random variable $Y_1, \ldots, Y_n$. In the case of sampling from a finite population under a design model, a renumbering of the selected units has been performed, of course, and the stochastic nature of $\mathbf{y}$ is determined by the sampling plan. The objective is to estimate some parameter θ. The observed data is denoted by $y_{obs} = \{(y_i : i \in s_r), s_r\}$, being the observed part of $\mathbf{y}$ and the response sample s_r of size n_r.

Let $\hat{\theta}$ be the estimator based on the full sample data $\mathbf{y}$, with $Var(\hat{\theta})$ estimated by $\hat{V}(\mathbf{y})$. For $i \in s - s_r$ we impute by some method y_i^* and let $\mathbf{y}^*$ denote the complete data $(y_i : i \in s_r, y_i^* : i \in s - s_r)$. Based on $\mathbf{y}*$, we have $\hat{\theta}^* = \hat{\theta}(\mathbf{y}^*)$ and $\hat{V}^* = \hat{V}(\mathbf{y}^*)$.

Multiple imputation of m repeated imputations leads to m completed data-sets with m estimates $\hat{\theta}_i^*, i = 1, \ldots, m$, and related variance estimates $\hat{V}_i^*, i = 1, \ldots, m$. The combined estimate is given by $\bar{\theta}^* = \sum_{i=1}^m \hat{\theta}_i^*/m$. The within-imputation variance is defined as $\bar{V}^* = \sum_{i=1}^m \hat{V}_i^*/m$ and the between-imputation component is $B^* = \sum_{i=1}^m (\hat{\theta}_i^* - \bar{\theta}^*)^2/(m-1)$. The total estimated variance of $\bar{\theta}^*$ is then proposed to be

$$W = \bar{V}^* + \left(k + \frac{1}{m}\right) B^* \tag{1}$$

That is, we need to determine k such that

$$E(W) = Var(\bar{\theta}^*) \tag{2}$$

Rubin (1987) has shown that $k = 1$ can be used with proper imputations, which essentially means drawing imputed values from a posterior distribution in a Bayesian framework.

In general, one has to determine the terms in (2). One way to try and do this is to use double expectation, conditioning on y_{obs}, that is, $E(W) = E\{E(W|Y_{obs})\}$ and $Var(\bar{\theta}^*) = E\{Var(\bar{\theta}^*|Y_{obs})\} + Var\{E(\bar{\theta}^*|Y_{obs})\}$. Typically,

$$E(\bar{V}^*) \approx Var(\hat{\theta}) \tag{3}$$

and $E(B^*|y_{obs}) = Var(\hat{\theta}^*|y_{obs})$. Hence, approximately

$$E(W) = Var(\hat{\theta}) + \left(E(k) + \frac{1}{m}\right) EVar(\hat{\theta}^*|Y_{obs}) \tag{4}$$

Moreover, $Var(\bar{\theta}^*|y_{obs}) = Var(\hat{\theta}^*|y_{obs})/m$ and $E(\bar{\theta}^*|y_{obs}) = E(\hat{\theta}^*|y_{obs})$. This implies that $Var(\bar{\theta}^*) = m^{-1}E\{Var(\hat{\theta}^*|Y_{obs})\} + Var\{E(\hat{\theta}^*|Y_{obs})\}$. From (3) and (4), Equation (2) becomes $Var(\hat{\theta}) + E(k)EVar(\hat{\theta}^*|Y_{obs}) = Var\{E(\hat{\theta}^*|Y_{obs})\}$, which gives the following general expression:

$$E(k) = \frac{VarE(\hat{\theta}^*|Y_{obs}) - Var(\hat{\theta})}{EVar(\hat{\theta}^*|Y_{obs})} \tag{5}$$

For this to be of interest, k must be, at least approximately, determined independently of unknown parameters. In addition, one needs to check that (3) holds. To illustrate how (5) can be used we shall in the next section consider three special cases with random nonresponse.

3. Three Applications for Random Nonresponse

3.1. Estimating Population Average with Hot-deck Imputation

Consider a simple random sample from a finite population of size N, where the aim is to estimate the population average μ of some variable y. We shall assume completely random nonresponse. In the terminology of Rubin (1987) and Little and Rubin (2002), the missingness mechanism is said to be MCAR (missing completely at random). We note that MCAR means that the response indicators $R_1, \ldots, R_N$ are independent with the same response probability $p_r = P(R_i = 1)$. The imputation method is the hot-deck method, where y_i^* is drawn at random from y_{obs} with replacement, and the estimate is the sample mean. Let $\bar{y}_r$ be the observed sample mean and $\hat{\sigma}_r^2 = \frac{1}{n_r - 1}\sum_{i \in s_r}(y_i - \bar{y}_r)^2$ the observed sample variance. Then $\bar{Y}^*$ is the imputation-based sample mean for the completed sample, and the combined estimator is given by $\bar{\bar{Y}}^* = \sum_{i=1}^m \bar{Y}_i^*/m$. Let $\bar{Y}_s$ denote the sample mean based on a full sample. Then, $Var(\bar{Y}_s) = \sigma^2(\frac{1}{n} - \frac{1}{N})$, with $\sigma^2 = (N-1)^{-1}\sum_{i=1}^N (y_i - \mu)^2$ being the population variance. We have further that $E(\bar{Y}^*|y_{obs}) = \bar{y}_r$ and $Var(\bar{Y}^*|y_{obs}) = \{(n - n_r)/n^2\}\{(n_r - 1)/n_r\}\hat{\sigma}_r^2$ using that $E(Y_i^*|y_{obs}) = \bar{y}_r$ and $Var(Y_i^*|y_{obs}) = \hat{\sigma}_r^2(n_r - 1)/n_r$.

In this case, $\hat{V}^* = \hat{\sigma}_*^2(\frac{1}{n} - \frac{1}{N})$ where $\hat{\sigma}_*^2 = \frac{1}{n-1}\left(\sum_{s_r}(y_i - \bar{y}^*)^2 + \sum_{s - s_r}(y_i^* - \bar{y}^*)^2\right)$. It can be shown that $E(\hat{\sigma}_*^2|y_{obs}) = \hat{\sigma}_r^2\left(1 - \frac{1}{n_r}\right)\left(1 + \frac{n_r}{n(n-1)}\right) \approx \hat{\sigma}_r^2$ and (3) holds.

We find, from (5),

$$E(k) = \frac{Var(\bar{Y}_r) - \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right)}{E\left(\frac{n-n_r}{n^2} \cdot \frac{n_r - 1}{n_r}\right) E(\hat{\sigma}_r^2 | n_r)}$$

$$= \frac{\sigma^2\left(E\left(\frac{1}{n_r}\right) - \frac{1}{N}\right) - \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right)}{E\left(\frac{n-n_r}{n^2} \cdot \frac{n_r - 1}{n_r}\right)\sigma^2} \approx \frac{(1-p_r)/p_r}{1-p_r} = \frac{1}{p_r}$$

which is satisfied approximately, with $f = (n - n_r)/n$ being the rate of nonresponse, by letting $k = 1/(1-f)$.

3.2. *Estimating the Regression Coefficient in the Ratio Model with Residual Imputation*

We shall assume completely random nonresponse as in Section 3.1. We consider a ratio model, i.e., regression through the origin: $Y_i = \beta x_i + \varepsilon_i$, with $Var(\varepsilon_i) = \sigma^2 x_i$; $i = 1, \ldots, n$. It is assumed that *all* x_i's are known, also in the nonresponse sample. The full data estimator of β is given by $\hat{\beta} = \sum_{i=1}^n Y_i / \sum_{i=1}^n x_i$. The unbiased estimator of σ^2 is given by $\hat{\sigma}^2 = \sum_{i=1}^n \frac{1}{x_i}(y_i - \hat{\beta} x_i)^2/(n-1)$.

We shall consider residual regression imputation. Let $\hat{\beta}_r$ be the $\hat{\beta}$ - estimate based on observed sample s_r. Define the standardized residuals $e_i = (y_i - \hat{\beta}_r x_i)/\sqrt{x_i}$, for $i \in s_r$. For $i \in s - s_r$: draw the value of e_i^* at random, with replacement, from the set of observed residuals $e_i, i \in s_r$. The imputed y-value is given by $y_i^* = \hat{\beta}_r x_i + e_i^* \sqrt{x_i}$.

Let $X = \sum_{i=1}^n x_i, X_r = \sum_{i \in s_r} x_i$ and $X_{nr} = \sum_{i \in s - s_r} x_i = X - X_r$. All considerations from now on are conditional on n_r and X_r, and we aim to determine k directly from (5). The proportion of the x-total in the nonresponse group is denoted as $f_X = X_{nr}/X$.

We now have $\hat{\beta}^* = \left(\sum_{s_r} y_i + \sum_{s-s_r} y_i^*\right)/X$ and $\hat{\sigma}_*^2 = \frac{1}{n-1}\left(\sum_{s_r} \frac{1}{x_i}(y_i - \hat{\beta}^* x_i)^2 + \sum_{s-s_r} \frac{1}{x_i}(y_i^* - \hat{\beta}^* x_i)^2\right)$.

In order to determine k from (5) we need to check the validity of (3) and derive $EVar(\hat{\beta}^* | y_{obs})$, $VarE(\hat{\beta}^* | y_{obs})$ and $Var(\hat{\beta})$. We note that $Var(\hat{\beta}) = \sigma^2/X$. In Appendix A.1 it is shown that condition (3) holds for moderate and large n_r, and that

$$\mathrm{Var}E(\hat{\beta}^* | y_{obs}) = \frac{\sigma^2}{X_r} + \frac{(1-d_1) d_2 n_{nr} X_{nr}}{X^2} \cdot \frac{\sigma^2}{n_r} \tag{6}$$

$$EVar(\hat{\beta}^* | y_{obs}) = \frac{X_{nr}}{X^2} \cdot \frac{\sigma^2}{n_r}(n_r + d_1 - 2) \tag{7}$$

Here, $0 \le d_1, d_2 \le 1$. From (5), using (6) and (7), we find

$$k = \frac{n_r X^2 - n_r X \cdot X_r + (1-d_1) d_2 n_{nr} X_{nr} X_r}{X_r X_{nr}(n_r + d_1 - 2)} \approx \frac{X}{X_r} + (1-d_1)\, d_2 \frac{n_{nr}}{n_r}$$

We note that if all $x_i = 1$, then $d_1 = d_2 = 1$. Now, with $f_X = X_{nr}/X$ being the proportion of the x-total in the nonresponse group and $f = n_{nr}/n$ the rate of nonresponse, we finally get, since typically $(1 - d_1)d_2 \approx 0$,

$$k \approx \frac{1}{1 - f_X} + (1 - d_1)d_2 \frac{f}{1 - f} \approx \frac{1}{1 - f_X}$$

for usual x-values and nonresponse rates.

3.3. Estimating the Regression Coefficient in Simple Linear Regression with Residual Imputation

As in Sections 3.1 and 3.2 the nonresponse mechanism is assumed to be MCAR with $p_r = P(R_i = 1)$. The simple linear regression model is assumed: $Y_i = \alpha + \beta x_i + \varepsilon_i$, with $Var(\varepsilon_i) = \sigma^2; i = 1, \ldots, n$. All x_i's are assumed to be known. We may assume, that $\bar{x} = \sum_{i=1}^n x_i/n = 0$. Then the full data estimates are given by $\hat{\beta} = \sum_{i=1}^n x_i y_i / SS_x$, where $SS_x = \sum_{i=1}^n x_i^2$, and $\hat{\alpha} = \bar{y} = \sum_{i=1}^n y_i/n$. The unbiased estimator of σ^2 is given by $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$. Let $\hat{\alpha}_r, \hat{\beta}_r$ be the estimates based on the response sample, $\hat{\alpha}_r = \bar{y}_r - \hat{\beta}_r \bar{x}_r$ and $\hat{\beta}_r = \sum_{i \in s_r}(x_i - \bar{x}_r) y_i / SS_{x,r}$. Here, $\bar{y}_r = \sum_{i \in s_r} y_i/n_r$, $\bar{x}_r = \sum_{i \in s_r} x_i/n_r$ and $SS_{x,r} = \sum_{i \in s_r}(x_i - \bar{x}_r)^2$.

Simple residual imputation is defined as follows: The observed residuals are $e_j = (y_j - \hat{\alpha}_r - \hat{\beta}_r x_j)$, for $j \in s_r$. For $i \in s - s_r$: draw e_i^* at random, with replacement from $(e_j, j \in s_r)$. The imputed y- value is given by $y_i^* = \hat{\alpha}_r + \hat{\beta}_r x_i + e_i^*$.

The imputation based estimates are $\hat{\beta}^* = \left(\sum_{i \in s_r} x_i y_i + \sum_{i \in s - s_r} x_i y_i^*\right)/SS_x$, $\hat{\alpha}^* = (n_r \bar{y}_r + (n - n_r)\bar{y}_{nr}^*)/n$ where $\bar{y}_{nr}^* = \sum_{s - s_r} y_i^*/(n - n_r)$ and $\hat{\sigma}_*^2 = \frac{1}{n-2}\left\{\sum_{s_r}(y_i - \hat{\alpha}^* - \hat{\beta}^* x_i)^2 + \sum_{s - s_r}(y_i^* - \hat{\alpha}^* - \hat{\beta}^* x_i)^2\right\}$. It can be shown (see Appendix A.2 for a summary proof) that $E(\hat{\sigma}_*^2) = \sigma^2 E(\frac{n_r - 2}{n_r} \cdot \frac{n - 2f}{n - 2}) \approx \sigma^2$ where, as in Section 3.1, $f = (n - n_r)/n$. Since $Var(\hat{\beta}) = \sigma^2/SS_x$, (3) holds. It is readily seen that $E(\hat{\beta}^* | y_{obs}) = \hat{\beta}_r$ and $Var(\hat{\beta}^* | y_{obs}) = s_e^2 \cdot c_r / SS_x$, where $s_e^2 = \sum_{s_r} e_i^2/n_r$ and $c_r = \sum_{s - s_r} x_i^2 / SS_x \in \langle 0, 1 \rangle$. It can be shown that $E(s_e^2 | s_r) = \frac{n_r - 2}{n_r}\sigma^2$. Moreover, clearly $E(c_r) = 1 - p_r$ and $Var(\hat{\beta}_r | s_r) = \sigma^2/SS_{x,r}$. It follows, from (5), that

$$E(k) = \frac{E(1/SS_{x,r}) - 1/SS_x}{(1 - p_r)E\{(n_r - 2)/n_r\}/SS_x} \approx \frac{1/E(SS_{x,r}) - 1/SS_x}{(1 - p_r)/SS_x}$$

Using the fact that conditional on n_r, s_r is a simple random sample such that the response indicators are correlated with $\mathrm{Cov}(R_i, R_j) = -f(1 - f)/(n - 1)$, we find that $E(SS_{x,r}) = (p_r - \frac{1 - p_r}{n - 1})SS_x$. It follows that, approximately, $E(k) = \frac{1}{p_r - \frac{1}{n}} \approx 1/p_r$ and we can use $k = 1/(1 - f)$.

4. Multiple Imputation for Stratified Samples

4.1. Separate Combinations

One way to combine the m completed data sets is to do it separately for each stratum, i.e., determine a separate k for each stratum. The general setup is then as follows:

The sample s is divided into H sample strata, $s_1, \ldots, s_H$. Let $\boldsymbol{y}_h$ be the planned full data from subsample s_h of size n_h. It is assumed that $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_H$ are independent. The observed part of $\boldsymbol{y}_h$ is denoted by $y_{h,obs}$ with s_{hr} being the response sample from s_h of size n_{hr}. The estimator based on the full sample data is the sum of independent terms, $\hat{\theta} = \sum_{h=1}^{H} \hat{\theta}_h$ where $\hat{\theta}_h$ is based on the $\boldsymbol{y}_h$. $Var(\hat{\theta}) = \sum_{h=1}^{H} Var(\hat{\theta}_h)$ is estimated by $\hat{V}(\hat{\theta}) = \sum_{h=1}^{H} \hat{V}_h(y_h)$ where $\hat{V}_h(y_h)$ is the variance estimate of $\hat{\theta}_h$ based on $\boldsymbol{y}_h$. For $i \in s_h - s_{hr}$ we impute by some method y_i^* based on $y_{h,obs}$ and let $\mathbf{y}_h^*$ denote the complete data $(y_{h,obs}, y_i^*, i \in s_h - s_{hr})$. Based on $\mathbf{y}_h^*$, we have $\hat{\theta}_h^* = \hat{\theta}_h(\mathbf{y}_h^*)$ and $\hat{V}_h^* = \hat{V}_h(\mathbf{y}_h^*)$. Then the imputation based estimator is given by $\hat{\theta}^* = \sum_{h=1}^{H} \hat{\theta}_h^*$ and $\hat{V}^* = \sum_{h=1}^{H} \hat{V}_h^*$. Multiple imputation of m repeated imputations leads to m completed data sets with m estimates for each stratum h, $\hat{\theta}_{h,i}, i = 1, \ldots, m$ and related variance estimates $\hat{V}_{h,i}^*, i = 1, \ldots, m$. The total estimates and related variances are $\hat{\theta}_i^* = \sum_{h=1}^{H} \hat{\theta}_{h,i}^*$ and $\hat{V}_i^* = \sum_{h=1}^{H} \hat{V}_{h,i}^*$; for $i = 1, \ldots, m$. The combined estimate for stratum h is given by $\bar{\theta}_h^* = \sum_{i=1}^{m} \hat{\theta}_{h,i}^*/m$. The within-imputation variance for stratum h is $\bar{V}_h^* = \sum_{i=1}^{m} \hat{V}_{h,i}^*/m$ and the between-imputation component is given by $B_h^* = \sum_{i=1}^{m} (\hat{\theta}_{h,i}^* - \bar{\theta}_h^*)^2/(m-1)$. Following the same idea as in Section 2, Formula (1), the total estimated variance of $\bar{\theta}_h^*$ is then proposed to be $W_h = \bar{V}_h^* + (k_h + \frac{1}{m})B_h^*$. The combined total estimate is given by $\bar{\theta}^* = \sum_{i=1}^{m} \hat{\theta}_i^*/m = \sum_{h=1}^{H} \bar{\theta}_h^*$. It follows that the total estimated variance of $\bar{\theta}^*$ can be expressed as

$$W_{sep} = \sum_{h=1}^{H} W_h = \bar{V}^* + \sum_{h=1}^{H} \left(k_h + \frac{1}{m}\right) B_h^* \tag{8}$$

where $\bar{V}^* = \sum_{i=1}^{m} \hat{V}_i^*/m = \sum_{h=1}^{H} \bar{V}_h^*$. Provided (3) holds for each stratum h,

$$E(\bar{V}_h^*) \approx Var(\hat{\theta}_h) \tag{9}$$

we have from (5) that k_h must satisfy

$$E(k_h) = \frac{VarE(\hat{\theta}_h^* | Y_{h,obs}) - Var(\hat{\theta}_h)}{EVar(\hat{\theta}_h^* | Y_{h,obs})} \tag{10}$$

The combination Formula (8) is an alternative to the usual combination Formula (1), especially useful when we get simple expressions for k_h but not for k. The next section develops an expression for k in this situation.

4.2. An Overall Combination Formula

Now let W be given by (1). We shall determine the between-imputation factor k. Since $E(W) = E(W_{sep})$ we have

$$E\left\{\sum_{h=1}^{H} \left(k_h + \frac{1}{m}\right) B_h^*\right\} = E\left(k + \frac{1}{m}\right) B^* \tag{11}$$

Here, $B^* = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}_i^* - \bar{\theta}^*)^2 = \frac{1}{m-1} \sum_{i=1}^{m} \left\{\sum_h (\hat{\theta}_{h,i}^* - \bar{\theta}_h^*)\right\}^2$. Note that $E(B^* | y_{obs}) = E\left(\sum_{h=1}^{H} B_h^* | y_{obs}\right)$, since $E(B^* | y_{obs}) = Var(\hat{\theta}^* | y_{obs}) = \sum_{h=1}^{H} Var(\hat{\theta}_h^* | y_{obs})$ and $E(B_h^* | y_{obs}) = Var(\hat{\theta}_h^* | y_{obs})$.

Hence, the identity (11) becomes $E\{\sum_{h=1}^{H} k_h E(B_h^*|Y_{obs})\} = E\{kE(B^*|Y_{obs})\}$. This gives us the solution $k = \sum_{h=1}^{H} k_h E(B_h^*|y_{obs})/E(B^*|y_{obs})$ if we want to use the usual combination Formula (1) and hence

$$k = \frac{\sum_{h=1}^{H} k_h Var(\hat{\theta}_h^*|y_{obs})}{Var(\hat{\theta}^*|y_{obs})} = \sum_{h=1}^{H} k_h \cdot \frac{Var(\hat{\theta}_h^*|y_{obs})}{Var(\hat{\theta}^*|y_{obs})} \tag{12}$$

a weighted average of k_h. We get a simple expression for k only when all k_h are equal, say $k_h = k_0$. Then $k = k_0$.

5. Four Applications to Stratified Samples and Random Nonresponse within Strata

5.1. Estimating Population Average from Stratified Sample with Stratified Hot-deck Imputation

Consider stratified simple random samples from a finite population of size N, with H strata of sizes N_h, $h = 1, \ldots, H$. The aim is to estimate the population average μ of some variable y. We assume completely random nonresponse within each stratum, denoted as MAR (missing at random) by Rubin (1987) and Little and Rubin (2002). This means that the response indicators in stratum h, $R_{h,1}, \ldots, R_{h,N_h}$ are independent with $p_{hr} = P(R_{h,i} = 1)$. The imputation method is stratified hot-deck. Let $y_{h,obs}$ be the observed part from the response sample s_{hr} of size n_{hr} from stratum h, $y_{h,obs} = (y_i : i \in s_{hr})$. Then an imputed value y_i^* in stratum h is drawn at random from $y_{h,obs}$. The estimator based on the full sample data is the usual stratified weighted average $\bar{Y}_{strat} = \sum_{h=1}^{H} N_h \bar{y}_h / N = \sum_{h=1}^{H} v_h \bar{y}_h$. Here, $v_h = N_h/N$ and $\bar{y}_h = \sum_{s_h} y_i/n_h$, where s_h is the sample from stratum h and $n_h = |s_h|$. Then $Var(\bar{Y}_{strat}) = \sum_{h=1}^{H} v_h^2 \sigma_h^2 (\frac{1}{n_h} - \frac{1}{N_h})$, with $\sigma_h^2 = \sum_{i \in U_h} (y_i - \mu_h)^2/(N_h - 1)$ being the population variance in stratum h. Here U_h is stratum population h and μ_h is the average in U_h.

Let $\bar{y}_{hr}$ be the observed sample mean from stratum h and $\hat{\sigma}_{hr}^2 = \frac{1}{n_{hr}-1} \sum_{i \in s_{hr}} (y_i - \bar{y}_{hr})^2$ the observed sample variance. The imputation-based estimator is given by $\bar{Y}_{strat}^* = \sum_{h=1}^{H} N_h \bar{y}_h^*/N$ where $\bar{y}_h^* = \left(\sum_{s_{hr}} y_i + \sum_{s_h - s_{hr}} y_i^*\right)/n_h = \left(n_{hr}\bar{y}_{hr} + \sum_{s_h - s_{hr}} y_i^*\right)/n_h$. Let the m imputation replicates of $\bar{Y}_{strat}^*$ be denoted by $\bar{Y}_{strat,i}^*$ for $i = 1, \ldots, m$. The combined estimator is given by $\bar{\bar{Y}}_{strat}^* = \sum_{i=1}^{m} \bar{Y}_{strat,i}^*/m$.

5.1.1. Separate Strata Combinations

It follows from Section 3.1 that $k_h = 1/(1 - f_h)$, where $f_h = (n_h - n_{hr})/n_h$ is the rate of nonresponse in stratum h. The combination formula for the variance estimate of $\bar{\bar{Y}}_{strat}^*$ becomes, from (8),

$$W_{sep} = \bar{V}^* + \sum_{h=1}^{H} \left(\frac{1}{1 - f_h} + \frac{1}{m}\right) B_h^*$$

Here, $\bar{V}^* = \sum_{h=1}^{H} \bar{V}_h^*$ and $\bar{V}_h^*$ is the average of the m values of the imputation-based variance estimate $\hat{V}_h^* = v_h^2 \hat{\sigma}_{h*}^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right)$ where $\hat{\sigma}_{h*}^2 = \frac{1}{n_h - 1} \left(\sum_{s_{hr}} (y_i - \bar{y}_h^*)^2 + \sum_{s_h - s_{hr}} (y_i^* - \bar{y}_h^*)^2\right)$.

5.1.2. Overall Combination Formula. Determination of k in (1)

From (12) we need to determine $Var(v_h\bar{Y}_h^*|y_{obs})$ and $Var(\bar{Y}_{strat}^*|y_{obs}) = \sum_{h=1}^{H} Var\ (v_h\bar{Y}_h^*|y_{obs})$. Then

$$k = \sum_{h=1}^{H} \frac{1}{1-f_h} \cdot \frac{Var(v_h\bar{Y}_h^*|y_{obs})}{Var(\bar{Y}_{strat}^*|y_{obs})}$$

Now, $E(\bar{Y}_h^*|y_{h,obs}) = \bar{y}_{hr}$ and $Var(\bar{Y}_h^*|y_{h,obs}) = \{(n_h - n_{hr})/n_h^2\} \cdot \{(n_{hr}-1)/n_{hr}\}\hat{\sigma}_{hr}^2 \approx f_h\hat{\sigma}_{hr}^2/n_h$. Hence we can determine k as

$$k = \sum_{h=1}^{H} \frac{1}{1-f_h} \cdot \frac{f_h v_h^2 \hat{\sigma}_{hr}^2/n_h}{\sum_{k=1}^{H} f_k v_k^2 \hat{\sigma}_{kr}^2/n_h}$$

If the stratum sizes N_h are large then we can let $\hat{V}(v_h\bar{Y}_h) = v_h^2\hat{\sigma}_{hr}^2/n_h$. Let also $b_h = f_h\hat{V}(v_h\bar{Y}_h)/\sum_{k=1}^{H} f_k\hat{V}(v_k\bar{Y}_k)$. Then

$$k = \frac{\sum_{h=1}^{H} \hat{V}(v_h\bar{Y}_h) f_h \frac{1}{1-f_h}}{\sum_{h=1}^{H} \hat{V}(v_h\bar{Y}_h) f_h} = \sum_{h=1}^{H} b_h \cdot \frac{1}{1-f_h} \tag{13}$$

Since $\sum_{h=1}^{H} b_h = 1$, we see that k is a weighted average of the inverse of the response rates. If all $f_h = f$, the overall nonresponse rate, we get as for simple random sample that $k = 1/(1-f)$. Otherwise, a stratum response rate $1 - f_h$ has large weight if either the nonresponse rate is large and/or the estimated variance of $v_h\bar{Y}_h$ is large.

5.1.3. An Alternative Expression for k in (1)

By directly applying (5) we can get an alternative expression for k. Given y_{obs}, the imputed sample means $\bar{Y}_h^*$ are independent, which implies that $E(\bar{Y}_{strat}^*|y_{obs}) = \sum_{h=1}^{H} N_h\bar{y}_{hr}/N = \bar{y}_{strat,r}$ and $Var(\bar{Y}_{strat}^*|y_{obs}) \approx \sum_{h=1}^{H} v_h^2 \cdot f_h\hat{\sigma}_{hr}^2/n_h$. Just like in Section 3.1, (3) holds. From (5) we get

$$E(k) \approx \frac{Var(\bar{Y}_{strat,r}) - Var(\bar{Y}_{strat})}{E\left(\sum_h v_h^2 \cdot f_h\hat{\sigma}_{hr}^2/n_h\right)}$$

$$= \frac{\sum_{h=1}^{H} v_h^2\sigma_h^2\left(E\left(\frac{1}{n_{hr}}\right) - \frac{1}{N_h}\right) - \sum_{h=1}^{H} v_h^2\sigma_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)}{\sum_{h=1}^{H} v_h^2 \cdot E\left\{\frac{f_h}{n_h}E(\hat{\sigma}_{hr}^2|n_{hr})\right\}}$$

$$\approx \frac{\sum_{h=1}^{H} v_h^2\sigma_h^2 \frac{1-p_{hr}}{n_h} \cdot \frac{1}{p_{hr}}}{\sum_{h=1}^{H} v_h^2\sigma_h^2 \frac{1-p_{hr}}{n_h}} = \frac{\sum_{h=1}^{H} v_h^2 \frac{\sigma_h^2}{n_{hr}} E(f_h) \frac{1-f_h}{E(1-f_h)}}{\sum_{h=1}^{H} v_h^2 \frac{\sigma_h^2}{n_{hr}} E(f_h)(1-f_h)} \tag{14}$$

Now, $Var(\bar{Y}_{hr}) = EVar(\bar{Y}_{hr}|n_{hr}) = \sigma_h^2 E(1/n_{hr})$. Let $\hat{V}(v_h\bar{Y}_{hr}) = v_h^2\hat{\sigma}_{hr}^2/n_{hr}$. Then we see that the expression for $E(k)$ is satisfied approximately, if the stratum sizes N_h are large, by letting

$$\frac{1}{k} = \frac{\sum_{h-1}^{H}(1-f_h)f_h\hat{V}(v_h\bar{Y}_{hr})}{\sum_{h-1}^{H}f_h\hat{V}(v_h\bar{Y}_{hr})} = \sum_{h-1}^{H}a_h(1-f_h) \tag{15}$$

where the weights $a_h = f_h\hat{V}(v_h\bar{Y}_{hr})/\sum_{k=1}^{H}f_k\hat{V}(v_k\bar{Y}_{kr})$. Since $\sum_{h=1}^{H}a_h = 1$, we see that $1/k$ is a weighted average of the response rates. If all $f_h = f$, the overall nonresponse rate, we have, as shown in Section 5.1.2, that $k = 1/(1-f)$. As seen in Section 5.1.2, we note also in Expression (15) that a stratum response rate $1-f_h$ has large weight if either the nonresponse rate is large and/or the estimated variance of $v_h\bar{Y}_{hr}$ is large. The estimate of the total based on the response sample is given by $\bar{Y}_{strat,r} = \sum_h v_h\bar{Y}_{hr}$. We obtain Formula (13) for k by noting from (14) that we have $E(k) \approx \sum_{h=1}^{H} Var(v_h\bar{Y}_h)E(f_h)\frac{1}{E(1-f_h)}/\sum_{h=1}^{H}Var(v_h\bar{Y}_h)E(f_h)$. Then we see that the expression for $E(k)$ is satisfied approximately, if the stratum sizes N_h are large, by letting k be given by (13).

5.2. Logistic Regression with Binary Explanatory Variable. Estimating Log(Odds Ratio)

The variables $Y_1, \ldots, Y_n$ are independent 0/1 -variables, and we have explanatory 0/1-variable x with fixed known values $x_1, \ldots, x_n$. The class probabilities are given by $\pi_1 = P(Y_i = 1|x_i = 1)$ and $\pi_0 = P(Y_i = 1|x_i = 0)$. We assume a MAR(missing at random) model for the response variables $R_1, \ldots, R_n$, with $P(R_i = 1|x_i = 1) = p_{1r}$ and $P(R_i = 1|x_i = 0) = p_{0r}$. We can reparametrize the model in a logit version, $\log\{P(Y = 1|x)/P(Y = 0|x)\} = \alpha + \beta x$, where $\alpha = \log\{\pi_0/(1-\pi_0)\}$ and $\beta = \log\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} =$ log(odds ratio). The aim is to estimate β. Let $s = (1, \ldots, n)$ denote the full sample with strata $s_1 = \{i \in s : x_i = 1\}$ and $s_0 = \{i \in s : x_i = 0\}$. The sizes of s_1 and s_0 are denoted by n_1 and n_0. We note that $n_1 = \sum_{i=1}^{n}x_i = X$ and $n_0 = n - X$. The response samples in the strata are $s_{1r} = \{i \in s_1 : R_i = 1\}$ and $s_{0r} = \{i \in s_0 : R_i = 1\}$ with total response sample being s_r of size n_r. Let also $n_{1r} = |s_{1r}|$ and $n_{0r} = |s_{0r}|$. We see that $n_{1r} = \sum_{s_r}x_i = X_r$ and $n_{0r} = n_r - X_r$. The data from s_r can be represented as follows where n_{ijr} denotes the number of observations with $x = i$ and $y = j$: see (Table 1).

We then have the maximum likelihood estimates (MLE) $\hat{\pi}_{1r} = n_{11r}/n_{1r}$ and $\hat{\pi}_{0r} = n_{01r}/n_{0r}$ and MLE of β equals $\hat{\beta}_r = \log\frac{\hat{\pi}_{1r}/(1-\hat{\pi}_{1r})}{\hat{\pi}_{0r}/(1-\hat{\pi}_{0r})} = \log(n_{11r}n_{00r}/n_{10r}n_{01r})$. Similarly, the

Table 1. The observed data and nonresponse totals for the two classes

$x\backslash y$	$y = 0$	$y = 1$	Totals	Nonresponse
$x = 0$	n_{00r}	n_{01r}	n_{0r}	$n_0 - n_{0r}$
$x = 1$	n_{10r}	n_{11r}	n_{1r}	$n_1 - n_{1r}$

estimator based on the full sample is given by $\hat{\beta} = \log \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_0/(1-\hat{\pi}_0)} = \log(n_{11}n_{00}/n_{10}n_{01})$ with obvious analogue notation. We can express this estimate as $\hat{\beta} = \log\{\hat{\pi}_1/(1-\hat{\pi}_1)\} - \log\{\hat{\pi}_0/(1-\hat{\pi}_0)\} = \hat{\beta}_1 - \hat{\beta}_0$, of the same form as in Section 4.1. We also have that $\hat{\beta}_1$ and $\hat{\beta}_0$ are independent based on the separate sample strata s_1 and s_0. For large n_0, n_1, $\hat{\beta}$ is approximately $N(\beta, \sigma^2_{\hat{\beta}})$ where $\sigma^2_{\hat{\beta}} = \{n_1\pi_1(1-\pi_1)\}^{-1} + \{n_0\pi_0(1-\pi_0)\}^{-1}$. So, approximately, $Var(\hat{\beta}_1) = 1/\{n_1\pi_1(1-\pi_1)\}$ and $Var(\hat{\beta}_0) = 1/\{n_0\pi_0(1-\pi_0)\}$ and an estimate of $Var(\hat{\beta})$ is given by

$$\hat{V}(\hat{\beta}) = \frac{1}{n_1\hat{\pi}_1(1-\hat{\pi}_1)} + \frac{1}{n_0\hat{\pi}_0(1-\hat{\pi}_0)} = \left(\frac{1}{n_{11}} + \frac{1}{n_{10}}\right) + \left(\frac{1}{n_{01}} + \frac{1}{n_{00}}\right)$$

such that $\hat{V}(\hat{\beta}) = \hat{V}_1 + \hat{V}_0$, where $\hat{V}_1 = \left(\frac{1}{n_{11}} + \frac{1}{n_{10}}\right)$ and $\hat{V}_0 = \left(\frac{1}{n_{01}} + \frac{1}{n_{00}}\right)$ are the variance estimates of $\hat{\beta}_1$ and $\hat{\beta}_0$, respectively.

We shall consider the following imputation method: For each missing value in $s_1 - s_{1r}$, the imputed value $y*$ is drawn at random from the estimated distribution of Y given $x = 1$:

$$y* = 1 \text{ with probability } \hat{\pi}_{1r} = n_{11r}/n_{1r} \text{ and } y* = 0 \text{ with probability } 1 - \hat{\pi}_{1r}.$$

The same imputation method is used for $s_0 - s_{0r}$ with $y*$ drawn at random from the estimated distribution of Y given $x = 0$. This is the same as stratified hot-deck imputation, imputed values are drawn at random, with replacement, from $y_{1,obs} = (y_i : i \in s_{1r})$ and $y_{0,obs} = (y_i : i \in s_{0r})$.

The imputed values in $s - s_r$ can be represented in the same form as the original data where now n^*_{ij} denotes the number of imputed values with $x = i$ and $y = j$: see (Table 2).

The imputation-based estimate of π_1 is given by $\hat{\pi}^*_1 = (n_{11r} + n^*_{11})/n_1$ such that the imputation-based estimate $\hat{\beta}^*_1 = \log\{\hat{\pi}^*_1/(1-\hat{\pi}^*_1)\} = \log\{(n_{11r} + n^*_{11})/(n_1 - n_{11r} - n^*_{11})\}$. Similarly, the imputation-based estimates for β_0 and β are given by $\hat{\beta}^*_0 = \log\{(n_{01r} + n^*_{01})/(n_0 - n_{01r} - n^*_{01})\}$ and $\hat{\beta}^* = \hat{\beta}^*_1 - \hat{\beta}^*_0$.

The m repeated imputations lead to m estimates $\hat{\beta}^*_{1,i}, \hat{\beta}^*_{0,i}, \hat{\beta}^*_i$, for $i = 1, \ldots, m$. The combined estimate is given by $\bar{\beta}^* = \sum_{i=1}^m \hat{\beta}^*_i/m = \sum_{i=1}^m \hat{\beta}^*_{1,i}/m - \sum_{i=1}^m \hat{\beta}^*_{0,i}/m = \bar{\beta}^*_1 - \bar{\beta}^*_0$. The imputed variance estimate $\hat{V}^*$ for $\hat{\beta}$ is given by

$$\hat{V}^* = \frac{1}{n_{11r} + n^*_{11}} + \frac{1}{n_{10r} + n^*_{10}} + \frac{1}{n_{01r} + n^*_{01}} + \frac{1}{n_{00r} + n^*_{00}} \tag{16}$$

We see that $E(\hat{V}^*|y_{obs}) \approx \frac{1}{n_1\hat{\pi}_{1r}(1-\hat{\pi}_{1r})} + \frac{1}{n_0\hat{\pi}_{0r}(1-\hat{\pi}_{0r})}$ and (3) hold. We also note that (9) holds separately for each class.

Table 2. The imputed totals for the two classes

$x\backslash y$	$y = 0$	$y = 1$	Totals
$x = 0$	n^*_{00}	n^*_{01}	$n_0 - n_{0r}$
$x = 1$	n^*_{10}	n^*_{11}	$n_1 - n_{1r}$

5.2.1. Separate Classes Combination

Let us first use the approach in Section 4.1 and determine separate k_1, k_0 for the two classes. Consider first stratum $s_1 = \{i \in s : x_i = 1\}$. In Appendix A.3 it is shown that $E(\hat{\beta}_1^*|y_{1,obs}) \approx \hat{\beta}_{1r}$ and $Var(\hat{\beta}_1^*|y_{1,obs}) \approx f_1(1-f_1)\hat{V}(\hat{\beta}_{1r})$. From (10), we find approximately:

$$E(k_1) = \frac{Var(\hat{\beta}_{1r}) - Var(\hat{\beta}_1)}{E\{f_1(1-f_1)\hat{V}(\hat{\beta}_{1r})\}} = \frac{E\,Var(\hat{\beta}_{1r}|n_{1r}) - Var(\hat{\beta}_1)}{E\{f_1(1-f_1)E[\hat{V}(\hat{\beta}_{1r})|n_{1r}]\}}$$

$$\approx \frac{\dfrac{1}{\pi_1(1-\pi_1)}\left(E\left(\dfrac{1}{n_{1r}}\right) - \dfrac{1}{n_1}\right)}{Ef_1(1-f_1)\dfrac{1}{n_{1r}\pi_1(1-\pi_1)}} \approx \frac{(1-p_{1r})/p_{1r}}{1-p_{1r}} = \frac{1}{p_{1r}}$$

which is satisfied approximately by letting $k_1 = 1/(1-f_1)$. In exactly the same way, we find that $k_0 = 1/(1-f_0)$ where $f_0 = (n_0 - n_{0r})/n_0$ is the rate of nonresponse in stratum s_0. The between-imputation component for $\hat{\beta}_1^*$ is given by $B_1^* = \frac{1}{m-1}\sum_{i=1}^m(\hat{\beta}_{1,i}^* - \bar{\beta}_1^*)^2$ and likewise B_0^* is the between-imputation component for $\hat{\beta}_0^*$. Then an estimated variance of the combined imputation-based estimate $\bar{\beta}^*$ for β is given by, from (8),

$$W_{sep} = \bar{V}^* + \sum\nolimits_{x=0}^{1}\left(\frac{1}{1-f_x} + \frac{1}{m}\right)B_x^*$$

where $\bar{V}^*$ is the average of m replicates of the imputed variance estimate $\hat{V}^*$ given by (16).

5.2.2. Overall Combination Formula. Determination of k in (1)

Since $Var(\hat{\beta}_1^*|y_{1,obs}) = f_1(1-f_1)\hat{V}(\hat{\beta}_{1r})$ and $Var(\hat{\beta}_0^*|y_{0,obs}) = f_0(1-f_0)\hat{V}(\hat{\beta}_{0r})$, we have from (12)

$$k = \frac{1}{1-f_1}\cdot\frac{f_1(1-f_1)\hat{V}(\hat{\beta}_{1r})}{\sum_{x=0}^{1}f_x(1-f_x)\hat{V}(\hat{\beta}_{xr})} + \frac{1}{1-f_0}\cdot\frac{f_0(1-f_0)\hat{V}(\hat{\beta}_{0r})}{\sum_{x=0}^{1}f_x(1-f_x)\hat{V}(\hat{\beta}_{xr})} \tag{17}$$

$Var(\hat{\beta}_1) \approx (n_{1r}/n_1)\,Var(\hat{\beta}_{1r} \mid n_{1r}) = (1-f_1)\,Var(\hat{\beta}_{1r}|n_{1r})$. Similarly, $Var(\hat{\beta}_0) \approx (1-f_0)\,Var(\hat{\beta}_{0r}|n_{0r})$. We can therefore estimate the variance of the full sample estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ by $\hat{V}(\hat{\beta}_1) = (1-f_1)\hat{V}(\hat{\beta}_{1r})$ and $\hat{V}(\hat{\beta}_0) = (1-f_0)\hat{V}(\hat{\beta}_{0r})$, respectively. Then

$$k = \frac{1}{1-f_1}\cdot\frac{f_1\hat{V}(\hat{\beta}_1)}{\sum_{x=0}^{1}f_x\hat{V}(\hat{\beta}_x)} + \frac{1}{1-f_0}\cdot\frac{f_0\hat{V}(\hat{\beta}_0)}{\sum_{x=0}^{1}f_x\hat{V}(\hat{\beta}_{xr})} = \frac{1}{1-f_1}\cdot b_1 + \frac{1}{1-f_0}\cdot(1-b_1)$$

Just like in Section 5.1.2 we see that k is a weighted average of the inverse of the response rates. If all $f_h = f$, the overall nonresponse rate, we get that $k = 1/(1-f)$. Otherwise, a stratum response rate $1 - f_x$ has large weight if either the nonresponse rate is large and/or the estimated variance of $\hat{\beta}_x$ is large.

Alternatively, from (17), $1/k = \sum_{x=0}^{1}(1-f_x)f_x\hat{V}(\hat{\beta}_{xr})/\sum_{x=0}^{1}f_x\hat{V}(\hat{\beta}_{xr}) = \sum_{x=0}^{1} a_x(1-f_x)$, where the weights are $a_x = f_x\hat{V}(\hat{\beta}_{xr})/\{f_1\hat{V}(\hat{\beta}_{1r}) + f_0\hat{V}(\hat{\beta}_{0r})\}$. So we can alternatively express $1/k$ as a weighted average of the response rates.

If the aim is to estimate π_1 and π_0 we obtain, of course, $k = 1/(1-f_1)$ for π_1 and $k = 1/(1-f_0)$ for π_0.

5.3. *Logistic Regression with Categorical Explanatory Variable. Estimating Log(Odds Ratios)*

If the explanatory x is categorical defining, say, H classes, we can generalize the results as follows:

Let $\pi_h = P(Y=1|x=h)$, $h = 0, \ldots, H-1$. Logistic regression defining the categories is done by introducing $H-1$ binary explanatory variables $x_1, \ldots, x_{H-1}$ where $x_h = 1$ if observation belongs to Class h, and 0 otherwise for $h = 1, \ldots, H-1$. Then an observation belongs to Class 0 if $x_1 = x_2 = \ldots = x_{H-1} = 0$. The logit version of the model becomes, with $\mathbf{x} = (x_1, x_2, \ldots, x_{H-1})$: $\log\{P(Y=1|\mathbf{x})/P(Y=0|\mathbf{x})\} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + x_{H-1}\beta_{H-1}$. We see that $\alpha = \log\frac{\pi_0}{1-\pi_0}$ and $\beta_h = \log\frac{\pi_h/(1-\pi_h)}{\pi_0/(1-\pi_0)} =$ log (odds ratio) for Class h versus Class 0. Estimating β_h by multiple imputation is done in exactly the same manner as for binary x, with Class h replacing Class 1.

5.4. *Logistic Regression with Missing Values in a Binary Explanatory Variable*

The situation is as in Section 5.2, except that y is fully observed in s, $\mathbf{y} = (y_1, \ldots, y_n)$, and we have missing values for the x-variable. $Y_1, \ldots, Y_n$ are independent 0/1-variables and we have an explanatory 0/1-variable x with fixed values $x_1, \ldots, x_n$, some of which are missing. The response variables indicate missingness of the x_i's but now with MAR model $P(R_i = 1|y_i = 1) = q_{1r}$ and $P(R_i = 1|y_i = 0) = q_{0r}$.

Otherwise, the model is the same as in Section 5.2 with class probabilities: $\pi_1 = P(Y_i = 1|x_i = 1)$ and $\pi_0 = P(Y_i = 1|x_i = 0)$, and the logit version $\log\{P(Y=1|x)/P(Y=0|x)\} = \alpha + \beta x$ with $\beta = \log\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$. The aim is still to estimate β.

Let now $s^1 = \{i \in s : y_i = 1\}$ and $s^0 = \{i \in s : y_i = 0\}$ with sizes n_1° and n_0°. The response samples in the strata are $s_r^1 = \{i \in s^1 : R_i = 1\}$ and $s_r^0 = \{i \in s^0 : R_i = 1\}$ with total response sample being $s_r = \{i \in s : R_i = 1\} = s_r^1 \cup s_r^0$. The data can now be represented as before, except that nonresponse totals are for each y-stratum. See Table 3.

The MLE $\hat{\pi}_{1r}, \hat{\pi}_{0r}, \hat{\beta}_r$, based on s_r are the same as before, as is the full sample estimate $\hat{\beta}$. The imputation method is stratified hot-deck for the y-strata. For each

Table 3. The observed data and nonresponse totals for the y-strata

$x\backslash y$	$y=0$	$y=1$
$x=0$	n_{00r}	n_{01r}
$x=1$	n_{10r}	n_{11r}
Totals	n_{0r}°	n_{1r}°
Nonresponse	$n_0^\circ - n_{0r}^\circ$	$n_1^\circ - n_{1r}^\circ$

missing value of x in $s^1 - s_r^1$, the imputed value $x*$ is drawn at random from $x_{1,obs} = (x_i : i \in s_r^1)$. Similarly, imputed values in $s^0 - s_r^0$ are drawn at random from $x_{0,obs} = (x_i : i \in s_r^0)$. The imputed values in $s - s_r$ can be represented in the same form as the original data where now n_{ij}^* denotes the number of imputed values with $x = i$ and $y = j$. See Table 4.

We need to find an approximate expression for the expectation and variance of $\hat{\beta}^*$, now denoted $\hat{\beta}_*$, conditional on the observed data. We defer to Appendix A.4 to show that $Var(\hat{\beta}_*|\boldsymbol{y}, x_{obs}) \approx f^1(1-f^1)(\frac{1}{n_{11r}} + \frac{1}{n_{01r}}) + f^0(1-f^0)(\frac{1}{n_{10r}} + \frac{1}{n_{00r}})$ and $E(\hat{\beta}_*|\boldsymbol{y}, x_{obs}) \approx \hat{\beta}_r$.

Here $f^1 = (n_1^\circ - n_{1r}^\circ)/n_1^\circ$ is the nonresponse rate in Stratum s^1 and $f^0 = (n_0^\circ - n_{0r}^\circ)/n_0^\circ$ the nonresponse rate in s^0. We note that $\hat{q}_{1r} = n_{1r}^\circ/n_1^\circ = 1 - f^1$ and $\hat{q}_{0r} = n_{0r}^\circ/n_0^\circ$. So the denominator in (5) becomes

$$E\left\{ f^1\left(1-f^1\right)\left(\frac{1}{n_{11r}} + \frac{1}{n_{01r}}\right) + f^0\left(1-f^0\right)\left(\frac{1}{n_{10r}} + \frac{1}{n_{00r}}\right)\right\} \tag{18}$$

The numerator in (5) equals, as before, $Var(\hat{\beta}_r) - Var(\hat{\beta})$, and we have approximately

$$Var(\hat{\beta}_r) - Var(\hat{\beta}) = \frac{1}{n_1\pi_1(1-\pi_1)} \cdot \frac{1-p_{1r}}{p_{1r}} + \frac{1}{n_0\pi_0(1-\pi_0)} \cdot \frac{1-p_{0r}}{p_{0r}} \tag{19}$$

where, as before, $p_{1r} = P(R_i = 1|x_i = 1)$ and $p_{0r} = P(R_i = 1|x_i = 0)$. We need alternative estimates of p_{1r} and p_{0r}. Since $p_{1r} = \pi_1 q_{1r} + (1-\pi_1)q_{0r}$, we have $\hat{p}_{1r} = \hat{\pi}_1(1-f^1) + (1-\hat{\pi}_1)(1-f^0)$. Similarly, $\hat{p}_{0r} = \hat{\pi}_0(1-f^1) + (1-\hat{\pi}_0)(1-f^0)$.

We can also use that $n_1\hat{p}_{1r} \approx n_{1r}$ and $n_0\hat{p}_{0r} \approx n_{0r}$. From (18) and (19) it follows that we can use

$$k = \frac{\left(\frac{1}{n_{11r}} + \frac{1}{n_{10r}}\right)\left(\hat{\pi}_{1r}f^1 + (1-\hat{\pi}_{1r})f^0\right) + \left(\frac{1}{n_{01r}} + \frac{1}{n_{00r}}\right)\left(\hat{\pi}_{0r}f^1 + (1-\hat{\pi}_{0r})f^0\right)}{f^1\left(1-f^1\right)\left(\frac{1}{n_{11r}} + \frac{1}{n_{01r}}\right) + f^0\left(1-f^0\right)\left(\frac{1}{n_{01r}} + \frac{1}{n_{00r}}\right)}$$

$$= \frac{f^1\left(\frac{1}{n_{10r}} + \frac{1}{n_{00r}}\right) + f^0\left(\frac{1}{n_{11r}} + \frac{1}{n_{01r}}\right)}{f^1\left(1-f^1\right)\left(\frac{1}{n_{11r}} + \frac{1}{n_{01r}}\right) + f^0\left(1-f^0\right)\left(\frac{1}{n_{01r}} + \frac{1}{n_{00r}}\right)}$$

We note that if $f^1 = f^0 = f$, then $k = 1/(1-f)$. Otherwise, we can express $1/k$ as a linear combination of the response rates $(1-f^1,\ 1-f^0)$. Let $w_1 = \frac{1}{n_{11r}} + \frac{1}{n_{01r}}$ and $w_0 = \frac{1}{n_{10r}} + \frac{1}{n_{00r}}$. Then

Table 4. The imputed totals for the y-strata

$x\backslash y$	$y = 0$	$y = 1$
$x = 0$	n_{00}^*	n_{01}^*
$x = 1$	n_{10}^*	n_{11}^*
Totals	$n_0^\circ - n_{0r}^\circ$	$n_1^\circ - n_{1r}^\circ$

$$\frac{1}{k} = a_1(1 - f^1) + a_0(1 - f^0)$$

where $a_1 = f^1 w_1/(f^1 w_0 + f^0 w_1)$ and $a_0 = f^0 w_0/(f^1 w_0 + f^0 w_1)$. We note that in general $a_1 + a_0 \neq 1$.

6. Question: Can We Use the Same Combination Formula for a Given Situation and Imputation Method for All Scientific Estimands?

We try here to give a general approach to this problem. Let s denote the full sample and $\mathbf{y}$ the full sample data. There are three possible cases:

1. s is a sample from a finite population and $\mathbf{y} = (y_i : i \in s)$ with design model. Then the observed stochastic variables are (s, s_r) and y_{obs} is equivalent to (s, s_r).
2. Same situation as in Case 1, but with a population model. Here, the observed stochastic variables are $y_{obs} = \{(y_i : i \in s_r), s_r, s\}$.
3. An observational study where $s = (1, \ldots, n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ is modeled. Then the observed stochastic variables are $y_{obs} = \{(y_i : i \in s_r), s_r\}$.

As an illustration we consider the case with nonresponse MCAR (the response variables R_i are independent with $p_r = P(R_i = 1)$) and hot-deck imputation. The case in Section 3.1 with a simple random sample is a special Case 1, and we found that for estimating the population mean with the sample mean,

$$k = \frac{1}{1 - f}, \text{ with } f = (n - n_r)/n = 1 - \hat{p}_r, \text{ the nonresponse rate} \tag{20}$$

Restricting attention to linear estimates where the imputed estimator $\hat{\theta}^*$ estimates the same parameter as $\hat{\theta}$, we will show that (20) holds in general for all the three cases above when the nonresponse mechanism is MCAR and we have hot-deck imputation. First, however, we consider the question whether hot-deck imputation always gives valid imputation-based estimators such that this value of k can be used. The answer, in general, is NO. One obvious requirement for an imputation method is that, at least approximately,

$$E(\hat{\theta}^* | \mathbf{y}, s) = \hat{\theta} \tag{21}$$

the imputed estimator should estimate the same parameter as $\hat{\theta}$. That is to say that conditional on the full planned data, the expected value of the imputed estimator should equal the full sample estimate. In Case 1, $\mathbf{y}$ is superfluous when s is given and (21) says that $E(\hat{\theta}^* | s) = \hat{\theta}$. In Case 3, s is not stochastic and therefore unnecessary, while in Case 2 we need both $\mathbf{y}$ and s.

We consider estimates that are linear in $(y_i : i \in s)$. The following results, proved in Appendix A.5, characterize linear estimates satisfying (21) with hot-deck imputation and show that for such estimators, $k = 1/(1 - f)$.

Lemma. Assume $\hat{\theta} = \sum_{i \in s} a_i(s) y_i$. Then $E(\hat{\theta}^* | \mathbf{y}, s) = \hat{\theta}$ if and only if $a_i(s) = a(s)$ for all $i \in s$.

That is $\hat{\theta} = a(s) \sum_{i \in s} y_i = n a(s) \bar{y}_s$.

Remark. In Case 3, s carries no information and $a_i(s) = a_i$.

Theorem. Consider $\hat{\theta} = \sum_{i \in s} a_i(s) y_i$ and $E(\hat{\theta}^* | \mathbf{y}, s) = \hat{\theta}$. Assume (3) holds. Then $E(k) = \frac{E(1/\hat{p}_r) - 1}{1 - p_r - \frac{1}{n}\{E(1/\hat{p}_r) - 1\}} \approx \frac{1}{p_r}$ and $k = 1/(1 - f)$ can be applied.

Let us look at some special cases:

1. With $a(s) = 1/n$, same as in Section 3.1, we see that (21) holds.
2. Regression coefficient for regression through the origin, $\hat{\beta} = \sum_{i=1}^{n} y_i / \sum_{i=1}^{n} x_i$. Here (21) is satisfied with $a = 1/\sum_{i=1}^{n} x_i$, and hence $k = 1/(1 - f)$.
3. A case where (21) does not hold is estimating the regression coefficient in usual linear regression where $\hat{\beta} = \sum (x_i - \bar{x}) y_i / \sum (x_i - \bar{x})^2$. Here, $a_i = (x_i - \bar{x}) / \sum_{j=1}^{n} (x_j - \bar{x})^2$, not independent of i. One can show that $E(\hat{\beta}^* | \mathbf{y}) \approx p_r \hat{\beta}$ (exact $\frac{np_r - 1}{n - 1} \hat{\beta}$). Hence, for regular regression problems hot-deck imputation cannot work. We note that from Section 3.3 one can use $k = 1/(1 - f)$ with residual imputation.

Obviously, when y is correlated to known x in a nonresponse group, one should utilize this in the imputations regardless of the estimation problems under consideration.

7. Discussion

We have shown that it is possible to develop a general theory for multiple imputation that does not require that the imputations are random draws from a Bayesian posterior distribution. For stratified samples with stratified estimates there is a need for further studies on which variance estimate to apply, either (8) using separate stratum combinations given by (10) or the overall combination (1) with k given by (12).

We see from the cases presented in this article that the non-Bayesian MI formula depends typically on a measure of the proportion of missing information in the response sample as compared to the full sample. In the simplest case in Section 3.1 the missing information is measured by $1/(1 - f)$, the inverse of the response rate. The higher this factor is, the more weight on the between-imputation component. In the ratio model in Section 3.2 with residual hot-deck regression imputation, the measure of missing information is the inverse of the proportion of the x-total in the response sample compared to the full sample. We note that in simple linear regression with the variance term independent of the explanatory variable, the missing information is again measured by $1/(1 - f)$. A suggestion for further study is to examine the possibility of generalizing this result by defining relevant measures of missing information, using the basic defining formula (5) for determining k.

It also remains to study the performance of related confidence intervals. Some preliminary simulation studies not included in this article show that for simple linear regression with residual imputation and $k = 1/(1 - f)$, confidence intervals of the form $\bar{\beta}^* \pm z_{\alpha/2} \sqrt{W}$ (where $z_{\alpha/2}$ is the upper $\alpha/2$-point in the N(0,1)-distribution) achieve approximately the nominal level $(1 - \alpha)$.

A. Appendix

A.1. Multiple Imputation for the Ratio Model in Section 3.2

Consider first the Condition (3) which is equivalent to $E(\hat{\sigma}_*^2) \approx \sigma^2$. Let $\hat{\beta}_{nr} = \sum_{s - s_r} y_i^* / X_{nr}$, and $\hat{\sigma}_{nr}^2 = \sum_{s = s_r} \frac{1}{x_i} (y_i^* - \hat{\beta}_{nr} x_i)^2 / (n_{nr} - 1)$. Here, $n_{nr} = n - n_r$.

Then one can express $\hat{\sigma}_*^2$ in the following way:

$$\hat{\sigma}_*^2 = \frac{1}{n-1}\left((n_r-1)\hat{\sigma}_r^2 + (n_{nr}-1)\hat{\sigma}_{nr}^2 + \frac{X_r X_{nr}}{X}(\hat{\beta}_r - \hat{\beta}_{nr})^2\right)$$

In this case, $E(Y_i^*|y_{obs}) = \hat{\beta}_r x_i + \bar{e}\sqrt{x_i}$, where $\bar{e} = \sum_{s_r} e_i/n_r$, and $Var(Y_i^*|y_{obs}) = x_i s_e^2$, where $s_e^2 = \frac{1}{n_r}\sum_{s_r}(e_i - \bar{e})^2$. Using this, it can be shown that

$$E(\hat{\sigma}_*^2) = \sigma^2\left(1 - \frac{c_1}{n-1} - \frac{4c_2}{(n-1)n_r} - c_3 f\frac{n-1}{n\cdot n_r}\right)$$

where c_1, c_2, c_3 lie in the interval (0,1). Hence, $E(\hat{\sigma}_*^2) \approx \sigma^2$ and (3) holds for moderate and large n_r.

Next, we look at $Var(\hat{\beta}^*|y_{obs})$ and $E(\hat{\beta}^*|y_{obs})$. We see that $\hat{\beta}^* = (\hat{\beta}_r X_r + \hat{\beta}_{nr}X_{nr})/X$, and $E(\hat{\beta}_{nr}|y_{obs}) = \hat{\beta}_r + (\bar{e}/X_{nr})\sum_{s-s_r}\sqrt{x_i}$ and $Var(\hat{\beta}_{nr}|y_{obs}) = s_e^2/X_{nr}$. This gives us $E(\hat{\beta}^*|y_{obs}) = \hat{\beta}_r + (\bar{e}/X)\sum_{s-s_r}\sqrt{x_i}$ and $Var(\hat{\beta}^*|y_{obs}) = (X_{nr}/X^2)s_e^2$. It follows that

$$VarE(\hat{\beta}^*|y_{obs}) = Var(\hat{\beta}_r) + \frac{\left(\sum_{s-s_r}\sqrt{x_i}\right)^2}{X^2}Var(\bar{e}) + 2\frac{\sum_{s-s_r}\sqrt{x_i}}{X}Cov(\hat{\beta}_r, \bar{e})$$

Now, $Cov(\hat{\beta}_r, \bar{e}) = 0$. Using Cauchy-Schwarz inequality, $\left(\sum a_i b_i\right)^2 \le \sum a_i^2 \sum b_i^2$ with $a_i = \sqrt{x_i}$ and $b_i = 1$, we see that $\left(\sum_{i=1}^n \sqrt{x_i}\right)^2 \le nX$. It follows that $Var(\bar{e}) = (\sigma^2/n_r)\left(1 - \left(\sum_{s_r}\sqrt{x_i}\right)^2/n_r X_r\right) = (1-d_1)\sigma^2/n_r$, $0 \le d_1 \le 1$, and $\left(\sum_{s-s_r}\sqrt{x_i}\right)^2/X^2 = d_2 n_{nr}X_{nr}/X^2, 0 \le d_2 \le 1$. Hence,

$$VarE(\hat{\beta}^*|y_{obs}) = \frac{\sigma^2}{X_r} + \frac{(1-d_1)d_2 n_{nr}X_{nr}}{X^2}\cdot\frac{\sigma^2}{n_r}$$

Next we find that $E(s_e^2) = \sigma^2(1 - \frac{1}{n_r}) - Var(\bar{e}) = \sigma^2(n_r + d_1 - 2)/n_r$, which gives us

$$EVar(\hat{\beta}^*|y_{obs}) = \frac{X_{nr}}{X^2}\cdot\frac{\sigma^2}{n_r}(n_r + d_1 - 2)$$

A.2. Multiple Imputation in Simple Linear Regression in Section 3.3. A Summary Proof of:

$$E(\hat{\sigma}_*^2) = \sigma^2 E\left(\frac{n_r - 2}{n_r}\cdot\frac{n-2f}{n-2}\right), \text{ where } f = (n-n_r)/n$$

$\hat{\sigma}_*^2 = \frac{1}{n-2}(SS_e^r + SS_e^{nr})$ where $SS_e^r = \sum_{s_r}(y_i - \hat{\alpha}^* - \hat{\beta}^* x_i)^2$ and $SS_e^{nr} = \sum_{s-s_r}(y_i^* - \hat{\alpha}^* - \hat{\beta}^* x_i)^2$
We can express the two residual sums of squares on the form

$$SS_e^r = \sum_{s_r}(y_i - \hat{\alpha}_r - \hat{\beta}_r x_i)^2 + \sum_{s_r}[(\hat{\alpha}_r - \hat{\alpha}^*)^2 + (\hat{\beta}_r - \hat{\beta}^*)x_i]^2$$

$$SS_e^{nr} = \sum_{s-s_r}(y_i - \hat{\alpha}_{nr}^* - \hat{\beta}_{nr}^* x_i)^2 + \sum_{s-s_r}[(\hat{\alpha}_{nr}^* - \hat{\alpha}^*) + (\hat{\beta}_{nr}^* - \hat{\beta}^*)x_i]^2$$

Here $\hat{\alpha}^*_{nr}, \hat{\beta}^*_{nr}$ are the estimates based only on the imputed $y^*_i, i \in s - s_r$. It follows that

$$E(SS^r_e|y_{obs}) = n_r s_e^2 + n_r Var(\hat{\alpha}^*|y_{obs}) + \left(\sum_{s-s_r} x_i^2\right) Var(\hat{\beta}^*|y_{obs}) + 2n_r\bar{x}_r Cov(\hat{\alpha}^*, \hat{\beta}^*|y_{obs})$$

$$= n_r s_e^2 + f(1-f)s_e^2 + (1-c_r)SS_x c_r s_e^2/SS_x + 2n_r\bar{x}_r f\bar{x}_{nr}s_e^2/SS_x$$

$$\left(\text{with } \bar{x}_{nr} = \sum_{s-s_r} x_i/(n-n_r)\right)$$

$$\Rightarrow E(SS^r_e|y_{obs}) = s_e^2\{n_r + f(1-f) + c_r(1-c_r) + 2n_r\bar{x}_r\bar{x}_{nr}f(1/SS_x)\} \quad (22)$$

After some algebraic manipulations, we find that

$$E(SS^{nr}_e|y_{obs}) = \sum_{s-s_r} Var(Y^*_i - \bar{Y}^*_{nr}|y_{obs}) + \sum_{s-s_r}(x_i - \bar{x}_{nr})^2 Var(\hat{\beta}^*_{nr}|y_{obs})$$

$$- 2\sum_{s-s_r}(x_i - \bar{x}_{nr})Cov(Y^*_i - \bar{Y}^*_{nr}, \hat{\beta}^*_{nr}|y_{obs})$$

$$+ (n-n_r)Var(\hat{\alpha}^*_{nr} - \hat{\alpha}^*|y_{obs}) + c_r SS_x Var(\hat{\beta}^*_{nr} - \hat{\beta}^*|y_{obs})$$

$$+ 2(n-n_r)\bar{x}_{nr}Cov(\hat{\alpha}^*_{nr} - \hat{\alpha}^*, \hat{\beta}^*_{nr} - \hat{\beta}^*|y_{obs})$$

$$= (n-n_r-1)s_e^2 + s_e^2 - 2s_e^2 + (n-n_r)\left\{(1-f)^2 s_e^2 \frac{1}{n-n_r} + \bar{x}^2_{nr}s_e^2\frac{1}{SS_{x,nr}}\right\}$$

$$+ c_r SS_x s_e^2\left(\frac{1}{SS_{x,nr}} + \frac{c_r}{SS_x} - 2\frac{1}{SS_x}\right) + 2(n-n_r)\bar{x}_{nr}\left(f\bar{x}_{nr}s_e^2\frac{1}{SS_x} - \bar{x}_{nr}s_e^2\frac{1}{SS_{x,nr}}\right)$$

where $SS_{x,nr} = \sum_{s-s_r}(x_i - \bar{x}_{nr})^2$. We see that $SS_{x,nr} = \sum_{s-s_r} x_i^2 - (n-n_r)\bar{x}^2_{nr} = c_r SS_x - (n-n_r)\bar{x}^2_{nr}$, and therefore $c_r SS_x/SS_{x,nr} = 1 + (n-n_r)\bar{x}^2_{nr}/SS_{x,nr}$. It follows that

$$E(SS^{nr}_e|y_{obs}) = s_e^2\left((n-n_r-1) + (1-f)^2 + c_r^2 - 2c_r + 2f\bar{x}^2_{nr}(n-n_r)/SS_x\right) \quad (23)$$

From (22) and (23) we find that

$$(n-2)E(\hat{\sigma}^2_*|y_{obs}) = s_e^2\left(n - f - c_r + 2f\frac{1}{SS_x}\bar{x}_{nr}n\bar{x}\right) = s_e^2(n-f-c_r)$$

Since $E(c_r|n_r) = \frac{1}{SS_x}\sum_i^n E(1-R_i|n_r)x_i^2 = (1-n_r/n) = f$, we have

$$(n-2)E(\hat{\sigma}^2_*) = Es_e^2(n-f-c_r) = E(n-f-c_r)E(s_e^2|s_r) = E(n-f-c_r)\frac{n_r-2}{n_r}\sigma^2$$

$$= \sigma^2 E\left\{\frac{n_r-2}{n_r}(n-f-E(c_r|n_r)\right\} = \sigma^2 E\left\{\frac{n_r-2}{n_r}(n-2f)\right\}$$

A.3. Logistic Regression with Binary Explanatory Variable. Separate Classes Combination

We shall determine $E(\hat{\beta}^*_1|y_{1,obs})$ and $Var(\hat{\beta}^*_1|y_{1,obs})$.

Conditional on $y_{1,obs,}$ n^*_{11} is binomially distributed $(n_1 - n_{1r}, \hat{\pi}_{1r})$. Hence, $E(n^*_{11}|y_{1,obs}) = (n_1 - n_{1r})\hat{\pi}_{1r}$ and $Var(n^*_{11}|y_{1,obs}) = (n_1 - n_{1r})\hat{\pi}_{1r}(1 - \hat{\pi}_{1r})$.

Conditional on $y_{1,obs}$, $\hat{\beta}_1^*$ is of the form $T = \log\{(a+Z)/(b-Z)\}$, where Z is binomial (n,p) and a and b are constants. Taylor linearization around $E(Z) = np$ gives $T \approx \log\{(a+np)/(b-np)\} + (Z-np)(a+b)/\{(a+z)(b-z)\}$ and

$$E(T) \approx \log\frac{a+np}{b-np} \quad \text{and } Var(T) \approx \left(\frac{a+b}{(a+np)(b-np)}\right)^2 np(1-p) \tag{24}$$

It follows that, with $a = n_{11r}$ and $b = n_1 - n_{11r}$, $E(\hat{\beta}_1^*|y_{1,obs}) \approx \hat{\beta}_{1r}$ and $Var(\hat{\beta}_1^*|y_{1,obs}) \approx \left(n_1/\{n_1\hat{\pi}_{1r}n_1(1-\hat{\pi}_{1r})\}\right)^2(n_1-n_{1r})\hat{\pi}_{1r}(1-\hat{\pi}_{1r})$. Let $f_1 = (n_1 - n_{1r})/n_1$ be the non-response rate in stratum s_1. We see that

$$Var(\hat{\beta}_1^*|y_{1,obs}) \approx \frac{f_1 n_1}{n_1^2}\cdot\frac{1}{\hat{\pi}_{1r}(1-\hat{\pi}_{1r})} = f_1(1-f_1)\cdot\frac{1}{n_{1r}\hat{\pi}_{1r}(1-\hat{\pi}_{1r})}$$
$$= f_1(1-f_1)\hat{V}(\hat{\beta}_{1r})$$

A.4. Logistic Regression With Missing Values in a Binary Explanatory Variable

To determine $Var(\hat{\beta}_*|\boldsymbol{y}, x_{obs})$ and $E(\hat{\beta}_*|\boldsymbol{y}, x_{obs})$ we need to represent $\hat{\beta}_*$ in a different way than in Section 5.2 for it to be the sum of two independent terms, conditional on the observed data $(\boldsymbol{y}, x_{obs})$:

$$\hat{\beta}_* = \log\frac{(n_{11r}+n_{11}^*)(n_{00r}+n_{00}^*)}{(n_{10r}+n_{10}^*)(n_{01r}+n_{01}^*)} = \log\frac{(n_{11r}+n_{11}^*)}{(n_{01r}+n_{01}^*)} - \log\frac{(n_{10r}+n_{10}^*)}{(n_{00r}+n_{00}^*)} = \hat{\beta}_*^1 - \hat{\beta}_*^0$$

and $Var(\hat{\beta}_*|\boldsymbol{y}, x_{obs}) = Var(\hat{\beta}_*^1|\boldsymbol{y}, x_{1,obs}) + Var(\hat{\beta}_*^0|\boldsymbol{y}, x_{0,obs})$. Conditional on $(\boldsymbol{y}, x_{obs})$, n_{11}^* is binomial $(n_1^\circ - n_{1r}^\circ, p^1)$ where $p^1 = n_{11r}/n_{1r}^\circ$, and n_{10}^* is binomial $(n_0^\circ - n_{0r}^\circ, p^0)$ where $p^0 = n_{10r}/n_{0r}^\circ$. Then, from (24), we find that approximately $E(\hat{\beta}_*^1|\boldsymbol{y}, x_{1,obs}) = \log\{p^1/(1-p^1)\}$ and $Var(\hat{\beta}_*^1|\boldsymbol{y}, x_{1,obs}) = \left(n_1^\circ/\{n_1^\circ p^1 n_1^\circ(1-p^1)\}\right)^2(n_1^\circ - n_{1r}^\circ)p^1(1-p^1)$. Then $Var(\hat{\beta}_*^1|\boldsymbol{y}, x_{1,obs}) \approx f^1/\{n_1^\circ p^1(1-p^1)\} = f^1(1-f^1)/\{n_{1r}^\circ p^1(1-p^1)\}$. Similarly, $E(\hat{\beta}_*^0|\boldsymbol{y}, x_{1,obs}) \approx \log\{p^0/(1-p^0)\}$ such that $E(\hat{\beta}_*|\boldsymbol{y}, x_{obs}) \approx \hat{\beta}_r$. Also, $Var(\hat{\beta}_*^0|\boldsymbol{y}, x_{0,obs}) \approx f^0/\{n_0^\circ p^0(1-p^0)\} = f^0(1-f^0)/\{n_{0r}^\circ p^0(1-p^0)\}$. We have that

$$\frac{1}{n_{1r}^\circ p^1(1-p^1)} = \frac{n_{1r}^\circ}{n_{11r}n_{01r}} = \frac{1}{n_{11r}} + \frac{1}{n_{01r}} \quad \text{and} \quad \frac{1}{n_{0r}^\circ p^0(1-p^0)} = \frac{1}{n_{10r}} + \frac{1}{n_{00r}}$$

and it follows that $Var(\hat{\beta}_*|\boldsymbol{y}, x_{obs}) \approx f^1(1-f^1)(\frac{1}{n_{11r}} + \frac{1}{n_{01r}}) + f^0(1-f^0)(\frac{1}{n_{10r}} + \frac{1}{n_{00r}})$.

A.5. Proofs of Lemma and Theorem in Section 6

In order to prove Lemma and Theorem in Section 6 we need some facts. In all three cases described in Section 6:

(a) n_r is binomial (n, p_r) and independent of s
(b) s_r given n_r,s is a simple random sample from s of size n_r
(c) $P(R_i = 1|n_r) = n_r/n$ and $P(R_i = 1, R_j = 1|n_r) = \frac{n_r}{n}\cdot\frac{n_r-1}{n-1}$ (follows from (b))
(d) $E(Y_i^*|y_{obs}) = \bar{y}_r$ ($\Rightarrow$ $E(Y_i^*|\boldsymbol{y}, s, n_r) = \bar{y}_s \Rightarrow E(Y_i^*|\boldsymbol{y}, s) = \bar{y}_s$)
(e) $Var(Y_i^*|y_{obs}) = \frac{n_r-1}{n_r}\hat{\sigma}_r^2$, where $\hat{\sigma}_r^2 = \frac{1}{n_r-1}\sum_{i\in s_r}(y_i - \bar{y}_r)^2$

(f) $E(\hat{\sigma}_r^2|\mathbf{y}, s, n_r) = \hat{\sigma}^2$ where $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i\in s}(y_i - \bar{y}_s)^2$
$(\Rightarrow Var(Y_i^*|\mathbf{y}, s, n_r) = \frac{n-1}{n}\hat{\sigma}^2 \approx \hat{\sigma}^2)$

(g) $Var(\bar{Y}_r|\mathbf{y}, s, n_r) = f\hat{\sigma}^2/n_r = \hat{\sigma}^2(\frac{1}{n_r} - \frac{1}{n})$

Proof of Lemma

$$E(\hat{\theta}^*|\mathbf{y}, s) = E\left\{ E\left(\sum_{i\in s_r} a_i(s)y_i + \sum_{i\in s-s_r} a_i(s)Y_i^*|Y_{obs} \right) \Big| \mathbf{y},\ s \right\}$$

$$=^{(d)} E\left(\sum_{i\in s_r} a_i(s)y_i|\mathbf{y}, s \right) + E\left(\sum_{i\in s-s_r} a_i(s)\bar{Y}_r|\mathbf{y}, s \right)$$

First term:

$$E\left(\sum_{i\in s_r} a_i(s)y_i|\mathbf{y}, s \right) = E\left\{ E\left(\sum_{i\in s} a_i(s)y_iR_i|\mathbf{y}, s, n_r \right)|\mathbf{y}, s \right\}$$

$$=^{(c)} E\left(\sum_{i\in s} a_i(s)y_i\frac{n_r}{n}|\mathbf{y}, s \right) =^{(a)} p_r\hat{\theta}$$

Second term:

$$E\left(\sum_{i\in s-s_r} a_i(s)\bar{Y}_r|\mathbf{y}, s \right) = E\left\{ E\left(\frac{1}{n_r}\sum_{i\in s}\sum_{j\in s} a_i(s)y_j(1-R_i)R_j|\mathbf{y}, s, n_r \right)|\mathbf{y}, s \right\}$$

$$=^{(c)} E\left(\frac{1}{n_r}\sum_{i\in s}\sum_{j\in s, j\neq i} a_i(s)y_j\left(\frac{n_r}{n} - \frac{n_r}{n}\cdot\frac{n_r-1}{n-1}\right)|\mathbf{y}, s \right)$$

$$=^{(a)} \frac{1-p_r}{n-1}\sum_{i\in s}\sum_{j\in s, j\neq i} a_i(s)y_j = \frac{1-p_r}{n-1}(n\bar{a}(s)n\bar{y}_s - \hat{\theta})$$

$$\text{where } \bar{a}(s) = \sum_{i\in s} a_i(s)/n$$

This implies that $E(\hat{\theta}^*|\mathbf{y}, s) = p_r\hat{\theta} + \frac{1-p_r}{n-1}(n^2\bar{a}(s)\bar{y}_s - \hat{\theta})$ and (21) $\Leftrightarrow \hat{\theta} = n\bar{a}(s)\bar{y}_s = \bar{a}(s)\sum_{i\in s} y_i$

Proof of Theorem

From Lemma, $\hat{\theta} = a(s)\sum_{i\in s} y_i = na(s)\bar{y}_s$ and $\hat{\theta}^* = a(s)\left(\sum_{i\in s_r} y_i + \sum_{i\in s-s_r} y_i^* \right)$

$E(\hat{\theta}^*|\, y_{obs}) =^{(d)} a(s)(n_r\bar{y}_r + (n-n_r)\bar{y}_r) = na(s)\bar{y}_r$

$Var(\hat{\theta}^*|\, y_{obs}) =^{(e)} \{a(s)\}^2(n-n_r)\frac{n_r-1}{n_r}\hat{\sigma}_r^2$

Hence,

$$\begin{aligned}VarE(\hat{\theta}^*|Y_{obs}) &= Var(na(s)\bar{Y}_r) = E\{n^2\{a(s)\}^2 Var(\bar{Y}_r|\mathbf{Y},s)\} + Var\{na(s)E(\bar{Y}_r|\mathbf{Y},s)\} \\ &= n^2 E\big([a(s)]^2\{E\{Var(\bar{Y}_r|Y,s,n_r)|\mathbf{Y},s\} + Var\{E(\bar{Y}_r|\mathbf{Y},s,n_r)|\mathbf{Y},s\}\big) \\ &\quad + Var\{na(s)E\{E(\bar{Y}_r|\mathbf{Y},s,n_r)|\mathbf{Y},s\}\} =^{(g)} n^2 E\big([a(s)]^2\{\hat{\sigma}^2 E\{(1/n_r) - 1/n)|s\} \\ &\quad + Var(\bar{Y}_s|\mathbf{Y},s\big) + Var\{na(s)\bar{Y}_s\} = nE\big([a(s)]^2\hat{\sigma}^2[E(1/\hat{p}_r|s) - 1] + 0\big) \\ &\quad + Var\hat{\theta} =^{(a)} n(E(1/\hat{p}_r) - 1)E\big([a(s)]^2\hat{\sigma}^2\big) + Var\hat{\theta}\end{aligned}$$

Next,

$$\begin{aligned}&EVar(\hat{\theta}^*|Y_{obs}) = E\left([a(s)]^2(n - n_r)\frac{n_r - 1}{n_r}\hat{\sigma}_r^2\right) \\ &= E\{(n - n_r)(1 - 1/n_r)[a(s)]^2 E(\hat{\sigma}_r^2|\mathbf{Y},s,n_r)\} = E\{(n - n_r)(1/n_r - 1)[a(s)]^2\hat{\sigma}^2\} \\ &=^{(a)} \{n(1 - p_r) - (E(1/\hat{p}_r) - 1)\}E([a(s)]^2\hat{\sigma}^2)\end{aligned}$$

We find now, from (5),

$$\begin{aligned}E(k) &= \frac{(E(1/\hat{p}_r) - 1)E\{[a(s)]^2\hat{\sigma}^2\}}{\{(1 - p_r) - \frac{1}{n}(E(1/\hat{p}_r) - 1)\}E\{[a(s)]^2\hat{\sigma}^2\}} = \frac{(E(1/\hat{p}_r) - 1)}{1 - p_r - \frac{1}{n}(E(1/\hat{p}_r) - 1)} \\ &\approx \frac{(1/p_r) - 1}{1 - p_r} = \frac{1}{p_r}.\end{aligned}$$

8. References

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New York: Wiley.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Rubin, D.B. (1996). Multiple Imputation after 18 + Years (with Discussion). Journal of the American Statistical Association, 91, 473–489.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

Received July 2005
Revised October 2005