Bjørnstad, Jan F.

Working Paper

Non-Bayesian Multiple Imputation

Discussion Papers, No. 421

Provided in Cooperation with:
Research Department, Statistics Norway, Oslo

Suggested Citation: Bjørnstad, Jan F. (2005) : Non-Bayesian Multiple Imputation, Discussion Papers, No. 421, Statistics Norway, Research Department, Oslo

This Version is available at:
https://hdl.handle.net/10419/192403

Discussion Papers No. 421, May 2005
Statistics Norway, Research Department

Jan F. Bjørnstad

Non-Bayesian Multiple Imputation

Abstract:
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 paper 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.

Keywords: Multiple imputation, survey sampling, nonresponse, hot-deck imputation

JEL classification: C42, C13, C15

Address: Jan F. Bjørnstad, Statistics Norway, Division for Statistical Methods and Standards, P.O. Box 8131 Dep., N-0033 Oslo, Norway. E-mail: jab@ssb.no

Discussion Papers	comprise research papers intended for international journals or books. A preprint of a Discussion Paper may be longer and more elaborate than a standard journal article, as it may include intermediate calculations and background material etc.

Abstracts with downloadable Discussion Papers
in PDF are available on the Internet:
http://www.ssb.no
http://ideas.repec.org/s/ssb/dispap.html

For printed Discussion Papers contact:

Statistics Norway
Sales- and subscription service
NO-2225 Kongsvinger

Telephone: +47 62 88 55 00
Telefax: +47 62 88 55 95
E-mail: Salg-abonnement@ssb.no

1. Introduction

Multiple imputation is a method specifically designed for variance estimation in the presence of missing data, developed by Rubin (1987). 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 being that the variability in nonproper imputations is too little 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 combination formula not dependent on unknown parameters. This paper 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 two applications with random nonresponse, (i) estimating a population average from simple random samples using hot-deck imputation and (ii) estimating a regression coefficient using 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 & response model.

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 $\mathbf{y} = (y_1,....,y_n)$ denoting the full sample data, values of random variable $Y_1,....,Y_n$. The objective is to estimate some parameter θ. Now, let y_{obs} be the observed part of **y**, with s_r being the response sample of size n_r,

$$y_{obs} = (y_i : i \in s_r).$$

Let $\hat{\theta}$ be the estimator based on the full sample data **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 **y*** denote the complete data $(y_{obs}, y_i^*, i \in s - s_r)$. Based on **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 $\hat{\theta}_i^*, i = 1,...,m,$ and related variance estimates $\hat{V}_i^*, i = 1,...,m$. 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

$$B^* = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}_i^* - \bar{\theta}^*)^2 .$$

The total estimated variance of $\bar{\theta}^*$ is then proposed to be

$$W = \bar{V}^* + (k + \frac{1}{m}) B^* . \qquad (1)$$

That is, we need to determine *k* such that

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

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,

$$E(W) = E\{E(W \mid Y_{obs})\}$$

$$Var(\bar{\theta}^*) = E\{Var(\bar{\theta}^* \mid Y_{obs})\} + Var\{E(\bar{\theta}^* \mid Y_{obs})\}.$$

Typically,

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

and

$$E(B^* \mid y_{obs}) = Var(\hat{\theta}^* \mid y_{obs}).$$

Hence, approximately

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

Moreover,

$$Var(\bar{\theta}^* \mid y_{obs}) = Var(\hat{\theta}^* \mid y_{obs})/m$$

and

$$E(\bar{\theta}^* \mid y_{obs}) = E(\hat{\theta}^* \mid y_{obs}).$$

This implies that

$$Var(\bar{\theta}^*) = \frac{1}{m}E\{Var(\hat{\theta}^* \mid Y_{obs})\} + Var\{E(\hat{\theta}^* \mid Y_{obs})\}.$$

From (3) and (4), the equation (2) becomes

$$Var(\hat{\theta}) + E(k)EVar(\hat{\theta}^* \mid Y_{obs}) = Var\{E(\hat{\theta}^* \mid Y_{obs})\},$$

which gives the following general expression for $E(k)$:

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

For this to be of interest, k must be, at least approximately, determined independent 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 two special cases with random nonresponse.

3. Two applications to simple random samples and random non-response

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 Rubin's term MCAR (missing completely at random). We note that MCAR means that the response indicators $R_1,...,R_N$ 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}, 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,

$$Var(\bar{Y}_s) = \sigma^2(\frac{1}{n} - \frac{1}{N}), \text{ with } \sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$$

being the population variance. We have further that

$$E(\bar{Y}^* \mid y_{obs}) = \bar{y}_r \quad \text{and } Var(\bar{Y}^* \mid y_{obs}) = \frac{n - n_r}{n^2} \cdot \frac{n_r - 1}{n_r}\hat{\sigma}_r^2$$

using that $E(Y_i^* \mid y_{obs}) = \bar{y}_r$ and $Var(Y_i^* \mid y_{obs}) = \frac{n_r - 1}{n_r}\hat{\sigma}_r^2$.

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

$$E(\hat{\sigma}_*^2 \mid y_{obs}) = \hat{\sigma}_r^2(1 - \tfrac{1}{n_r})(1 + \tfrac{n_r}{n(n-1)}) \approx \hat{\sigma}_r^2$$

and (3) holds. We find, from (5),

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

$$= \frac{\sigma^2(E(\frac{1}{n_r}) - \frac{1}{N}) - \sigma^2(\frac{1}{n} - \frac{1}{N})}{E(\frac{n-n_r}{n^2} \cdot \frac{n_r - 1}{n_r})\sigma^2}$$

$$\approx \frac{(1-p_r)/p_r}{1-p_r} = \frac{1}{p_r}$$

which is satisfied approximately by letting

$$k = \frac{1}{1-f}$$

where $f = (n - n_r)/n$ is the rate of nonresponse.

3.2. Estimating regression coefficient 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:

$$Y_i = \beta x_i + \varepsilon_i \text{ , with } Var(\varepsilon_i) = \sigma^2 x_i \text{ ; } i = 1, \ldots, n.$$

It is assumed that <u>all</u> x_i's are known, also in the nonresponse sample. The full data estimator of β is given by

$$\hat{\beta} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} x_i \text{ .}$$

The unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \frac{1}{x_i} (y_i - \hat{\beta} x_i)^2 \text{ .}$$

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

$$e_i = (y_i - \hat{\beta}_r x_i)/\sqrt{x_i} \text{ , for } i \in s_r \text{ .}$$

For $i \in s - s_r$: Draw the value of e_i^* at random from the set of observed residuals $e_i, i \in s_r$, and the imputed *y*-value is given by

$$y_i^* = \hat{\beta}_r x_i + e_i^* \sqrt{x_i} \text{ .}$$

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). Define the proportion of the x-total in the nonresponse group to be:

$$f_X = X_{nr} / X \ .$$

We now have

$$\hat{\beta}^* = (\sum_{s_r} y_i + \sum_{s-s_r} y_i^*) / X$$

$$\hat{\sigma}_*^2 = \frac{1}{n-1}(\sum_{s_r} \tfrac{1}{x_i}(y_i - \hat{\beta}^* x_i)^2 + \sum_{s-s_r} \tfrac{1}{x_i}(y_i^* - \hat{\beta}^* x_i)^2) \ .$$

In order to determine k from (5) we need to check the validity of (3) and derive the following quantities: $Var(\hat{\beta}^* \mid y_{obs}), E(\hat{\beta}^* \mid y_{obs})$ and $Var(\hat{\beta})$. We note that

$$Var(\hat{\beta}) = \sigma^2 / X \ .$$

Consider (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 $\hat{\sigma}_{nr}^2 = \frac{1}{n_{nr}-1} \sum_{s-s_r} \frac{1}{x_i}(y_i^* - \hat{\beta}_{nr} x_i)^2$. Here, $n_{nr} = n - n_r$. Then, after some algebra, 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,

$$E(Y_i^* \mid y_{obs}) = \hat{\beta}_r x_i + \bar{e}\sqrt{x_i} \ , \text{ where } \bar{e} = \sum_{s_r} e_i / n_r \ ,$$

$$Var(Y_i^* \mid y_{obs}) = x_i s_e^2 \ , \text{ where } s_e^2 = \tfrac{1}{n_r} \sum_{s_r} (e_i - \bar{e})^2 \ .$$

Using this, it can be shown that

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

where c_1, c_2, c_3 lies in the interval (0,1).

Hence, $E(\hat{\sigma}_*^2) \approx \sigma^2$ and (3) follows, at least for moderate and large n_r.

Next, we look at $Var(\hat{\beta}^* \mid y_{obs})$ and $E(\hat{\beta}^* \mid y_{obs})$:

We see that $\hat{\beta}^* = (\hat{\beta}_r X_r + \hat{\beta}_{nr} X_{nr})/X$, and

$$E(\hat{\beta}_{nr} \mid y_{obs}) = \hat{\beta}_r + \frac{\bar{e}}{X_{nr}} \sum_{s-s_r} \sqrt{x_i}$$

$$Var(\hat{\beta}_{nr} \mid y_{obs}) = s_e^2 / X_{nr}.$$

This gives us

$$E(\hat{\beta}^* \mid y_{obs}) = \hat{\beta}_r + \frac{\bar{e}}{X} \sum_{s-s_r} \sqrt{x_i}$$

$$Var(\hat{\beta}^* \mid y_{obs}) = \frac{X_{nr}}{X^2} s_e^2.$$

Next, we need to find $EVar(\hat{\beta}^* \mid y_{obs})$ and $\mathrm{Var}E(\hat{\beta}^* \mid y_{obs})$:

$$\mathrm{Var}E(\hat{\beta}^* \mid y_{obs}) = Var(\hat{\beta}_r) + \frac{(\sum_{s-s_r} \sqrt{x_i})^2}{X^2} Var(\bar{e}) + 2\frac{\sum_{s-s_r} \sqrt{x_i}}{X} Cov(\hat{\beta}_r, \bar{e}).$$

Using Cauchy-Schwarz inequality,

$$(\sum_{i=1}^{n} a_i b_i)^2 \le \sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2$$

with $a_i = \sqrt{x_i}$ and $b_i = 1$, we see that

$$(\sum_{i=1}^{n} \sqrt{x_i})^2 \le nX. \tag{6}$$

Now, after some algebra we find that $Cov(\hat{\beta}_r, \bar{e}) = 0$ and

$$Var(\bar{e}) = \frac{\sigma^2}{n_r}\left(1 - \frac{(\sum_{s_r} \sqrt{x_i})^2}{n_r X_r}\right) = (1 - d_1)\frac{\sigma^2}{n_r},\ 0 \le d_1 \le 1.$$

Moreover, from (6),

$$\frac{(\sum_{s-s_r} \sqrt{x_i})^2}{X^2} = \frac{d_2 n_{nr} X_{nr}}{X^2},\quad 0 \le d_2 \le 1.$$

Hence,

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

Next we find that

$$E(s_e^2) = \sigma^2(1-\tfrac{1}{n_r}) - Var(\bar{e}) = \frac{\sigma^2}{n_r}(n_r + d_1 - 2)$$

which gives us

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

From (5),

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

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

$$\approx \frac{X}{X_r} + (1-d_1)d_2 \frac{n_{nr}}{n_r}.$$

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 goup 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.

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, that is determine k . The general setup is then as follows: The sample s is divided into H sample strata, $s_1, \ldots, s_H$. Let $\mathbf{y}_h$ be the planned full data from sub sample s_h of size n_h. It is assumed that $\mathbf{y}_1, \ldots, \mathbf{y}_H$ are independent. The observed part of $\mathbf{y}_h$ 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:

$$\hat{\theta} = \sum_{h=1}^{H} \hat{\theta}_h \quad \text{where } \hat{\theta}_h \text{ is based on the } \mathbf{y}_h.$$

$Var(\hat{\theta}) = \sum_{h=1}^{H} Var(\hat{\theta}_h)$ 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 $\mathbf{y}_h$. 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

$$B_h^* = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}_{h,i}^* - \bar{\theta}_h^*)^2 .$$

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

$$W_{sep} = \sum_{h=1}^{H} W_h = \bar{V}^* + \sum_{h=1}^{H} (k_h + \frac{1}{m}) B_h^* \tag{7}$$

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*,

$$E(\bar{V}_h^*) \approx Var(\hat{\theta}_h) \tag{8}$$

we have from (5) that k_h must satisfy

$$E(k_h) = \frac{VarE(\hat{\theta}_h^* \mid Y_{h,obs}) - Var(\hat{\theta}_h)}{EVar(\hat{\theta}_h^* \mid Y_{h,obs})} . \tag{9}$$

The combination formula (7) 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 developes 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

$$E\{\sum_{h=1}^{H} (k_h + \frac{1}{m}) B_h^*\} = E(k + \frac{1}{m}) B^* . \tag{10}$$

Here, $B^* = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}_i^* - \bar{\theta}^*)^2 = \frac{1}{m-1} \sum_{i=1}^{m} \{\sum_h (\hat{\theta}_{h,i}^* - \bar{\theta}_h^*)\}^2$. We note that

$$E(B^* \mid y_{obs}) = E(\textstyle\sum_{h=1}^{H} B_h^* \mid y_{obs}) .$$

This follows from the fact that $E(B^* \mid y_{obs}) = Var(\hat{\theta}^* \mid y_{obs}) = \sum_{h=1}^{H} Var(\hat{\theta}_h^* \mid y_{obs})$ and $E(B_h^* \mid y_{obs}) = Var(\hat{\theta}_h^* \mid y_{obs})$.

Hence, the identity (10) becomes

$$E\{\sum_{h=1}^{H} k_h E(B_h^* \mid Y_{obs})\} = E\{kE(B^* \mid Y_{obs})\}.$$

This gives us a solution for *k* if we want to use the usual combination formula (1):

$$k = \frac{\sum_{h=1}^{H} k_h E(B_h^* \mid y_{obs})}{E(B^* \mid y_{obs})}$$

$$= \frac{\sum_{h=1}^{H} k_h Var(\hat{\theta}_h^* \mid y_{obs})}{Var(\hat{\theta}^* \mid y_{obs})} = \sum_{h=1}^{H} k_h \cdot \frac{Var(\hat{\theta}_h^* \mid y_{obs})}{Var(\hat{\theta}^* \mid y_{obs})}, \qquad (11)$$

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, $h = 1,...,H$. The aim is to estimate the population average μ of some variable *y*. We assume completely random nonresponse within each stratum, typically denoted as MAR (missing at random). This means that the response indicators in stratum *h*, $R_{h,1},...,R_{h,N_h}$ are independent with the same response probability $p_{hr} = P(R_{h,i} = 1)$. 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

$$\bar{Y}_{strat} = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} v_h \bar{y}_h \ .$$

Here, $v_h = N_h / N$ and $\bar{y}_h = \sum_{i \in s_h} y_i / n_h$, where s_h is the sample from stratum h and $n_h = | s_h |$.

Then

$$Var(\bar{Y}_{strat}) = \sum_{h=1}^{H} v_h^2 \sigma_h^2 (\frac{1}{n_h} - \frac{1}{N_h}), \text{ with } \sigma_h^2 = \frac{1}{N_h - 1}\sum_{i \in U_h} (y_i - \mu_h)^2$$

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

$$\bar{Y}_{strat}^* = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_h^*$$

where

$$\bar{y}_h^* = \frac{1}{n_h}(\sum_{i \in s_{hr}} y_i + \sum_{i \in s_h - s_{hr}} y_i^*) = \frac{1}{n_h}(n_{hr}\bar{y}_{hr} + \sum_{i \in s_h - s_{hr}} y_i^*) .$$

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

$$\bar{\bar{Y}}_{strat}^* = \sum_{i=1}^{m} \bar{Y}_{strat,i}^* \ .$$

5.1.1. Separate strata combinations

It follows from Section 3.1 that

$$k_h = \frac{1}{1 - f_h}$$

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 (7),

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

Here, $\overline{V}^* = \sum_{h=1}^{H} \overline{V}_h^*$ and $\overline{V}_h^*$ is the average of the *m* values of the imputation based variance estimate

$$\hat{V}_h^* = v_h^2 \hat{\sigma}_{h*}^2 (\frac{1}{n_h} - \frac{1}{N_h})$$

where

$$\hat{\sigma}_{h*}^2 = \frac{1}{n_h - 1} \left(\sum_{s_{hr}} (y_i - \overline{y}_h^*)^2 + \sum_{s_h - s_{hr}} (y_i^* - \overline{y}_h^*)^2 \right).$$

5.1.2. Overall combination formula. Determination of k in (1)

From (11) we need to determine $Var(v_h \overline{Y}_h^* \mid y_{obs})$ and $Var(\overline{Y}_{strat}^* \mid y_{obs}) = \sum_{h=1}^{H} Var(v_h \overline{Y}_h^* \mid y_{obs})$. Then we have that

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

Now, for $i \in s_h - s_{hr}$:

$$E(Y_i^* \mid y_{h,obs}) = \overline{y}_{hr} \text{ and } Var(Y_i^* \mid y_{h,obs}) = \frac{n_{hr} - 1}{n_{hr}} \hat{\sigma}_{hr}^2.$$

This gives the following results:

$$E(\overline{Y}_h^* \mid y_{h,obs}) = \overline{y}_{hr} \text{ and } Var(\overline{Y}_h^* \mid y_{h,obs}) = \frac{n_h - n_{hr}}{n_h^2} \cdot \frac{n_{hr} - 1}{n_{hr}} \hat{\sigma}_{hr}^2 \approx f_h \frac{\hat{\sigma}_{hr}^2}{n_h}.$$

Hence we can determine *k* as

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

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

$$k = \frac{\sum_{h=1}^{H} \hat{V}(v_h \overline{Y}_h) f_h \frac{1}{1 - f_h}}{\sum_{h=1}^{H} \hat{V}(v_h \overline{Y}_h) f_h} = \sum_{h=1}^{H} b_h \cdot \frac{1}{1 - f_h}. \qquad (12)$$

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 $v_h\overline{Y}_h$ 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 $\overline{Y}_h^*$ are independent which implies that

$$E(\overline{Y}_{strat}^* \mid y_{obs}) = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_{hr} = \bar{y}_{strat,r} \text{ and } Var(\overline{Y}_{strat}^* \mid y_{obs}) = \sum_{h=1}^{H} v_h^2 \cdot \frac{n_h - n_{hr}}{n_h^2} \cdot \frac{n_{hr} - 1}{n_{hr}} \hat{\sigma}_{hr}^2 .$$

It follows that

$$Var(\overline{Y}_{strat}^* \mid y_{obs}) \approx \sum_{h=1}^{H} v_h^2 \cdot \frac{f_h}{n_h} \hat{\sigma}_{hr}^2 .$$

Just like in Section 3.1, (3) holds. From (5) we get

$$E(k) \approx \frac{Var(\overline{Y}_{strat,r}) - Var(\overline{Y}_{strat})}{E(\sum_{h=1}^{H} v_h^2 \cdot \frac{f_h}{n_h} \hat{\sigma}_{hr}^2)}$$

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

$$= \frac{\sum_{h=1}^{H} v_h^2 \sigma_h^2 [E(\frac{1}{n_{hr}}) - \frac{1}{n_h}]}{\sum_{h=1}^{H} v_h^2 \cdot \sigma_h^2 \frac{E(f_h)}{n_h}}$$

$$\approx \frac{\sum_{h=1}^{H} v_h^2 \sigma_h^2 \frac{1 - p_{hr}}{n_h} \cdot \frac{1}{p_{hr}}}{\sum_{h=1}^{H} v_h^2 \sigma_h^2 \frac{1 - p_{hr}}{n_h}} = \frac{\sum_{h=1}^{H} v_h^2 \frac{\sigma_h^2}{n_{hr}} E(f_h) \frac{1 - f_h}{E(1 - f_h)}}{\sum_{h=1}^{H} v_h^2 \frac{\sigma_h^2}{n_{hr}} E(f_h)(1 - f_h)} . \quad (13)$$

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

$$\frac{1}{k} = \frac{\sum_{h=1}^{H} (1 - f_h) f_h \hat{V}(v_h \overline{Y}_{hr})}{\sum_{h=1}^{H} f_h \hat{V}(v_h \overline{Y}_{hr})} = \sum_{h=1}^{H} a_h (1 - f_h) \quad (14)$$

where the weights $a_h = f_h\hat{V}(v_h\overline{Y}_{hr}) / \sum_{k=1}^{H} f_k\hat{V}(v_k\overline{Y}_{kr})$. 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 (14) that a stratum response rate 1- f_h has large weight if either the nonresponse rate is large and/or the estimated variance of $v_h\overline{Y}_{hr}$ is large. We note that the estimate of the total based on the response sample is given by

$$\overline{Y}_{strat,r} = \sum_{h=1}^{H} v_h\overline{Y}_{hr}.$$

We obtain formula (12) for *k* by noting from (13) that we can express *E*(*k*) as

$$E(k) \approx \frac{\sum_{h=1}^{H} Var(v_h\overline{Y}_h)E(f_h)\dfrac{1}{E(1-f_h)}}{\sum_{h=1}^{H} Var(v_h\overline{Y}_h)E(f_h)}.$$

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 (12).

5.2. Logistic regression with binary explanatory variable. Estimating log(odds ratio)

The model is as follows:

$Y_1,...,Y_n$ are independent 0/1 -variables

Explanatory 0/1-variable *x* with fixed known values $x_1,...,x_n$

Class probabilities: $\pi_1 = P(Y_i = 1 \mid x_i = 1)$ and $\pi_0 = P(Y_i = 1 \mid x_i = 0)$

Response variables: $R_1,...,R_n$ with MAR (missing at random) model:

$P(R_i = 1 \mid x_i = 1) = p_{1r}$ and $P(R_i = 1 \mid x_i = 0) = p_{0r}$

We can reparametrize the model in a logit version:

$$\log\frac{P(Y=1 \mid x)}{P(Y=0 \mid x)} = \alpha + \beta x$$

giving us the following 1-1 relationships:

$$\alpha = \log\frac{\pi_0}{1-\pi_0} \Leftrightarrow \pi_0 = \frac{1}{1+e^{-\alpha}}$$

$$\beta = \log\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \log(\text{odds ratio}), \text{ and } \pi_1 = \frac{1}{1+e^{-(\alpha+\beta)}}.$$

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 $y = j$:

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

We then have the (maximum likelihood) estimates MLE)

$$\hat{\pi}_{1r} = n_{11r} / n_{1r} \text{ and } \hat{\pi}_{0r} = n_{01r} / n_{0r}.$$

Hence, MLE of β equals

$$\hat{\beta}_r = \log \frac{\hat{\pi}_{1r} /(1 - \hat{\pi}_{1r})}{\hat{\pi}_{0r} /(1 - \hat{\pi}_{0r})} = \log \frac{n_{11r} n_{00r}}{n_{10r} n_{01r}}.$$

Similarily, the estimator based on the full sample is given by

$$\hat{\beta} = \log \frac{\hat{\pi}_1 /(1 - \hat{\pi}_1)}{\hat{\pi}_0 /(1 - \hat{\pi}_0)} = \log \frac{n_{11} n_{00}}{n_{10} n_{01}}$$

with obvious analogue notation. We can express this estimate as follows:

$$\hat{\beta} = \log \frac{\hat{\pi}_1 /(1 - \hat{\pi}_1)}{\hat{\pi}_0 /(1 - \hat{\pi}_0)} = \log \frac{\hat{\pi}_1}{1 - \hat{\pi}_1} - \log \frac{\hat{\pi}_0}{1 - \hat{\pi}_0} = \hat{\beta}_1 - \hat{\beta}_0,$$

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. It can be shown that for large n_0, n_1, $\hat{\beta}$ is approximately $N(\beta, \sigma^2_{\hat{\beta}})$ where

$$\sigma^2_{\hat{\beta}} = \frac{1}{n_1 \pi_1 (1 - \pi_1)} + \frac{1}{n_0 \pi_0 (1 - \pi_0)}.$$

Here, approximately, $Var(\hat{\beta}_1) = 1/\{n_1 \pi_1 (1 - \pi_1)\}$ and $Var(\hat{\beta}_0) = 1/\{n_0 \pi_0 (1 - \pi_0)\}$. It follows that an estimate of $Var(\hat{\beta})$ is given by

$$\hat{V}(\hat{\beta}) = \frac{1}{n_1 \hat{\pi}_1 (1 - \hat{\pi}_1)} + \frac{1}{n_0 \hat{\pi}_0 (1 - \hat{\pi}_0)}$$

$$= \frac{n_1}{n_{11} n_{10}} + \frac{n_0}{n_{01} n_{00}} = \left(\frac{1}{n_{11}} + \frac{1}{n_{10}} \right) + \left(\frac{1}{n_{01}} + \frac{1}{n_{00}} \right).$$

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

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 $x = 1$:

$y^* = 1$ with probability $\hat{\pi}_{1r} = n_{11r} / n_{1r}$ and $y^* = 0$ with probability $1 - \hat{\pi}_{1r} = n_{10r} / n_{1r}$.

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 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 $y = j$:

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

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 $\hat{\beta}_1^*$ becomes

$$\hat{\beta}_1^* = \log \frac{\hat{\pi}_1^*}{1 - \hat{\pi}_1^*} = \log \frac{n_{11r} + n_{11}^*}{n_1 - n_{11r} - n_{11}^*}.$$

Similarily, the imputation based estimates for β_0 and β are given by

$$\hat{\beta}_0^* = \log \frac{n_{01r} + n_{01}^*}{n_0 - n_{01r} - n_{01}^*} \text{ and } \hat{\beta}^* = \hat{\beta}_1^* - \hat{\beta}_0^*.$$

The m repeated imputations leads 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

$$\hat{V}^* = \frac{1}{n_{11r} + n_{11}^*} + \frac{1}{n_{10r} + n_{10}^*} + \frac{1}{n_{01r} + n_{01}^*} + \frac{1}{n_{00r} + n_{00}^*}. \quad (15)$$

We see that $E(\hat{V}^* \mid y_{obs}) \approx \frac{1}{n_1 \hat{\pi}_{1r}(1-\hat{\pi}_{1r})} + \frac{1}{n_0 \hat{\pi}_{0r}(1-\hat{\pi}_{0r})}$ and (3) holds. We also note that (8) holds separately for each class.

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 $s_1 = \{i \in s : x_i = 1\}$. To determine k_1 from (9) we need to determine $E(\hat{\beta}_1^* \mid y_{1,obs})$ and $Var(\hat{\beta}_1^* \mid y_{1,obs})$. We have that conditional on $y_{1,obs}$, n_{11}^* is binomially distributed $(n_1 - n_{1r}, \hat{\pi}_{1r})$. Hence,

$$E(n_{11}^* \mid y_{1,obs}) = (n_1 - n_{1r})\hat{\pi}_{1r} \text{ and } Var(n_{11}^* \mid y_{1,obs}) = (n_1 - n_{1r})\hat{\pi}_{1r}(1-\hat{\pi}_{1r}).$$

We see that, conditional on $y_{1,obs}$, $\hat{\beta}_1^*$ is of the form

$$T = \log \frac{a+Z}{b-Z}$$, where Z is binomial (n,p) and a and b are constants.

Taylor linearization around $E(Z) = np$ gives that

$$T \approx \log \frac{a+np}{b-np} + (Z-np)\frac{a+b}{(a+z)(b-z)}$$

and

$$E(T) \approx \log \frac{a+np}{b-np} \text{ and } Var(T) \approx \left(\frac{a+b}{(a+np)(b-np)} \right)^2 np(1-p). \qquad (16)$$

It follows that, with a = n_{11r} and $b = n_1$ - n_{11r} :

$$E(\hat{\beta}_1^* \mid y_{1,obs}) \approx \log \frac{n_{11r} + (n_1 - n_{1r})\hat{\pi}_{1r}}{n_1 - n_{11r} - (n_1 - n_{1r})\hat{\pi}_{1r}} = \log \frac{\hat{\pi}_{1r}}{1-\hat{\pi}_{1r}} = \hat{\beta}_{1r}$$

and

$$Var(\hat{\beta}_1^* \mid y_{1,obs}) \approx \left(\frac{n_1}{n_1 \hat{\pi}_{1r} n_1 (1-\hat{\pi}_{1r})} \right)^2 (n_1 - n_{1r})\hat{\pi}_{1r}(1-\hat{\pi}_{1r}).$$

Let f_1 be the nonresponse rate in stratum s_1: $f_1 = (n_1 - n_{1r})/n_1$. We see that

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

From (9), we find approximately:

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

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

$$\approx \frac{\frac{1}{\pi_1(1-\pi_1)}(E(\frac{1}{n_{1r}}) - \frac{1}{n_1})}{Ef_1(1-f_1)\frac{1}{n_{1r}\pi_1(1-\pi_1)}}$$

$$\approx \frac{\frac{1}{n_1 p_{1r}} - \frac{1}{n_1}}{\frac{1}{n_1}Ef_1} = \frac{(1-p_{1r})/p_{1r}}{1-p_{1r}} = \frac{1}{p_{1r}}$$

which is satisfied approximately by letting

$$k_1 = \frac{1}{1-f_1}.$$

In exactly the same way, we find that

$$k_0 = \frac{1}{1-f_0}$$

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 $B_1^* = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\beta}_{1,i}^* - \overline{\beta}_1^*)^2$ and likewise B_0^* is the between imputation component for $\hat{\beta}_0^*$. Then an estimated variance of the combined imputation based estimate $\overline{\beta}^*$ for β is given by, from (7),

$$W_{sep} = \overline{V}^* + \sum_{x=0}^{1}(\frac{1}{1-f_x} + \frac{1}{m})B_x^*$$

where $\overline{V}^*$ is the average of *m* replicates of the imputed variance estimate $\hat{V}^*$ given by (15).

5.2.2. Overall combination formula. Determination of k in (1)

From (11) we need $Var(\hat{\beta}_1^* \mid y_{1,obs})$ and $Var(\hat{\beta}_0^* \mid y_{0,obs})$. We have from previous section that

$$Var(\hat{\beta}_1^* \mid y_{1,obs}) = f_1(1-f_1)\hat{V}(\hat{\beta}_{1r})$$

$$Var(\hat{\beta}_0^* \mid y_{0,obs}) = f_0(1-f_0)\hat{V}(\hat{\beta}_{0r}).$$

It follows from (11) that

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

Now, $Var(\hat{\beta}_1) \approx (n_{1r}/n_1)Var(\hat{\beta}_{1r} \mid n_{1r}) = (1-f_1)Var(\hat{\beta}_{1r} \mid n_{1r})$. Similarily, $Var(\hat{\beta}_0) \approx (1-f_0)Var(\hat{\beta}_{0r} \mid n_{0r})$. 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):

$$\frac{1}{k} = \frac{\sum_{x=0}^{1}(1-f_x)f_x\hat{V}(\hat{\beta}_{xr})}{\sum_{x=0}^{1} f_x\hat{V}(\hat{\beta}_{xr})} = \sum_{x=0}^{1} a_x(1-f_x)$$

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.

We note that 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 $\pi_h = P(Y=1 \mid x=h)$, $h = 0, \ldots, H-1$. 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 $\mathbf{x} = (x_1, x_2, \ldots, x_{H-1})$:

$$\log\frac{P(Y=1 \mid \mathbf{x})}{P(Y=0 \mid \mathbf{x})} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + x_{H-1}\beta_{H-1}.$$

We see that

$$\alpha = \log \frac{\pi_0}{1-\pi_0}$$

and

$$\beta_h = \log \frac{\pi_h /(1-\pi_h)}{\pi_0 /(1-\pi_0)} = \log(\text{odds ratio}) \text{ for class } h \text{ versus class } 0.$$

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*, $\boldsymbol{y} = (y_1,...,y_n)$, and we have missing values for the *x*- variable. $Y_1,...,Y_n$ are independent 0/1 -variables and we have an explanatory 0/1-variable *x* with fixed values $x_1,...,x_n$, some of which are missing. The response variables indicate missingness of the x_i's with now with MAR model

$$P(R_i = 1 \mid y_i = 1) = q_{1r} \text{ and } P(R_i = 1 \mid y_i = 0) = q_{0r}.$$

Otherwise, the model is the same as in Section 5.2 with class probabilities: $\pi_1 = P(Y_i = 1 \mid x_i = 1)$ and $\pi_0 = P(Y_i = 1 \mid x_i = 0)$, and the logit version $\log\{P(Y = 1 \mid x) / P(Y = 0 \mid x)\} = \alpha + \beta x$ 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 is for each *y*-stratum.

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

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 missing value of x in $s^1 - s_r^1$, the imputed value x^* is drawn at random from $x_{1,obs} = (x_i : i \in s_r^1)$. Similarily, 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 $y = j$:

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

Now we need to represent $\hat{\beta}^*$, now denoted $\hat{\beta}_*$, in a different way for it to be the sum of two independent terms, conditional on the observed data ($\mathbf{y}$, 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

$$Var(\hat{\beta}_* \mid \mathbf{y}, x_{obs}) = Var(\hat{\beta}_*^1 \mid \mathbf{y}, x_{1,obs}) + Var(\hat{\beta}_*^0 \mid \mathbf{y}, x_{0,obs}).$$

We see that

$$\hat{\beta}_*^1 = \log \frac{n_{11r} + n_{11}^*}{n_1^\circ - n_{11r} - n_{11}^*} \quad \text{and} \quad \hat{\beta}_*^0 = \log \frac{n_{10r} + n_{10}^*}{n_0^\circ - n_{10r} - n_{10}^*}.$$

We now have that conditional on ($\mathbf{y}$, 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 (16), we find that approximately:

$$E(\hat{\beta}_*^1 \mid \mathbf{y}, x_{1,obs}) \approx \log \frac{n_{11r} + (n_1^\circ - n_{1r}^\circ)p^1}{n_1^\circ - n_{11r} - (n_1^\circ - n_{1r}^\circ)p^1} = \log \frac{n_1^\circ p^1}{n_1^\circ (1 - p^1)} = \log \frac{p^1}{(1 - p^1)}$$

and

$$Var(\hat{\beta}_*^1 \mid \boldsymbol{y}, x_{1,obs}) \approx \left(\frac{n_1^\circ}{n_1^\circ p^1 n_1^\circ (1-p^1)} \right)^2 (n_1^\circ - n_{1r}^\circ) p^1 (1-p^1) .$$

Let f^1 be the nonresponse rate in stratum s^1: $f^1 = (n_1^\circ - n_{1r}^\circ)/n_1^\circ$. We note that $\hat{q}_{1r} = n_{1r}^\circ / n_1^\circ = 1 - f^1$. We see that

$$Var(\hat{\beta}_*^1 \mid \boldsymbol{y}, x_{1,obs}) \approx f^1 \frac{1}{n_1^\circ p^1 (1-p^1)} = f^1 (1-f^1) \frac{1}{n_{1r}^\circ p^1 (1-p^1)} .$$

Similarily,

$$E(\hat{\beta}_*^0 \mid \boldsymbol{y}, x_{1,obs}) \approx \log \frac{p^0}{(1-p^0)}$$

$$Var(\hat{\beta}_*^0 \mid \boldsymbol{y}, x_{0,obs}) \approx f^0 \frac{1}{n_0^\circ p^0 (1-p^0)} = f^0 (1-f^0) \frac{1}{n_{0r}^\circ p^0 (1-p^0)}$$

where f^0 is the nonresponse rate in s^0: $f^0 = (n_0^\circ - n_{0r}^\circ)/n_0^\circ$. We have that

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

So the denominator in (5) becomes

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

To obtain the numerator in (5) we first see that:

$$E(\hat{\beta}_* \mid \boldsymbol{y}, x_{obs}) \approx \log \frac{p^1}{1-p^1} - \log \frac{p^0}{1-p^0}$$

$$= \log \frac{n_{11r}}{n_{01r}} - \log \frac{n_{10r}}{n_{00r}} = \log \frac{n_{11r} n_{00r}}{n_{01r} n_{10r}} = \hat{\beta}_r .$$

Hence, the numerator in (5) equals, as before, $Var(\hat{\beta}_r) - Var(\hat{\beta})$, and exactly as before 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

$$p_{1r} = P(R_i = 1 \mid x_i = 1) \text{ and } p_{0r} = P(R_i = 1 \mid x_i = 0) .$$

We need alternative estimates of p_{1r} and p_{0r}:

Since

$$p_{1r} = P(R_i = 1 \mid x_i = 1) = P(R_i = 1, Y_i = 1 \mid x_i = 1) + P(R_i = 1, Y_i = 0 \mid x_i = 1)$$
$$= P(Y_i = 1 \mid x_i = 1) P(R_i = 1 \mid Y_i = 1) + P(Y_i = 0 \mid x_i = 1) P(R_i = 1 \mid Y_i = 0)$$
$$= \pi_1 q_{1r} + (1 - \pi_1) q_{0r},$$

we have $\hat{p}_{1r} = \hat{\pi}_1 (1 - f^1) + (1 - \hat{\pi}_1)(1 - f^0)$ and $1 - \hat{p}_{1r} = \hat{\pi}_1 f^1 + (1 - \hat{\pi}_1) f^0$.

Similarily, $\hat{p}_{0r} = \hat{\pi}_0 (1 - f^1) + (1 - \hat{\pi}_0)(1 - f^0)$ and $1 - \hat{p}_{0r} = \hat{\pi}_0 f^1 + (1 - \hat{\pi}_0) f^0$.

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

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

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

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

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

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

where

$$a_1 = f^1 w_1 / (f^1 w_0 + f^0 w_1) \text{ and } a_0 = f^0 w_0 / (f^1 w_0 + f^0 w_1).$$

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. As an illustration we consider the case in Section 3.1, a simple random sample with nonresponse MCAR and hot-deck imputation. For other situations and imputation methods, similar considerations should be studied.

In Section 3.1 we found that for estimating the population mean with the sample mean,

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

The question is now: Is this *k* valid for *other* estimation problems as well, using the same imputation method. The answer, in general, is NO. What is needed is to find conditions for (20) to be valid. In this case, the stochastic variables are (s,s_r), so an alternative notation is to use (s,s_r) instead of Y_{obs} . Hence, (5) becomes

$$E(k) = \frac{VarE(\hat{\theta}^* \mid s, s_r) - Var(\hat{\theta})}{EVar(\hat{\theta}^* \mid s, s_r)}. \tag{21}$$

One obvious requirement is that, at least approximately

$$E(\hat{\theta}^* \mid s) = \hat{\theta}, \tag{22}$$

the imputed estimator should estimate the same parameter as $\hat{\theta}$.

We shall in this note restrict attention to estimates that are linear in $(y_i : i \in s)$:

$$\hat{\theta} = \sum_{i \in s} a_i(s) y_i \tag{23}$$

Some results:

<u>*Lemma 1*</u> Assume $\hat{\theta}$ is given by (23). Then $\hat{\theta}$ satisfies (22) if and only if $a_i(s) = a(s)$ for all $i \in s$.

I.e., $\hat{\theta} = a(s)\sum_{i \in s} y_i = na(s)\bar{y}_s$.

<u>*Theorem*</u> Assume $\hat{\theta}$ is given by (23) and satisfies (22). Then $E(k) = 1/p_r$ and $k = 1/(1-f)$.

Before we prove these two results, let us look at some special cases:

1. $a(s) = 1/n$, same as in Section 3.1.
2. Regression coefficient for regression through the origin, $\hat{\beta} = \sum_{i=s} y_i / \sum_{i \in s} x_i$. Here (22) is satisfied with $a(s) = 1/\sum_{i \in s} x_i$, and hence $k = 1/(1-f)$.

3. A case where (22) does not hold is estimating the regression coefficient in usual linear regression:

$$\hat{\beta} = \frac{\sum_{i\in s}(x_i - \bar{x}_s)y_i}{\sum_{i\in s}(x_i - \bar{x}_s)^2}.$$

Here, $a_i(s) = \dfrac{x_i - \bar{x}_s}{\sum_{j\in s}(x_j - \bar{x}_s)^2}$, not independent of *i*.

Here one can show that $E(\hat{\beta}^* \mid s) \approx p_r\hat{\beta}$ (exact $\frac{np_r - 1}{n-1}\hat{\beta}$). Hence, for regular regression problems hot-deck imputation cannot work.

Obviously, when *y* is correlated to known *x* in nonresponse group, one should utilize this in the imputations regardless of the estimation problems under consideration.
In order to prove the two results we need some facts:

- (a) n_r is binomial (n,p_r)
- (b) s_r given s,n_r is a simple random sample from s of size n_r
- (c) $P(R_i = 1 \mid s, n_r) = n_r / n$ and $P(R_i = 1, R_j = 1 \mid s, n_r) = \dfrac{n_r}{n}\cdot\dfrac{n_r - 1}{n-1}$ (follows from (b))
- (d) $E(Y_i^* \mid s, s_r) = \bar{y}_r$ ($\Rightarrow E(Y_i^* \mid s, n_r) = \bar{y}_s \Rightarrow E(Y_i^* \mid s) = \bar{y}_s$)
- (e) $Var(Y_i^* \mid s, s_r) = \dfrac{n_r - 1}{n_r}\hat{\sigma}_r^2$

($\Rightarrow Var(Y_i^* \mid s, n_r) = \dfrac{n_r - 1}{n_r}\hat{\sigma}_s^2$, where $\hat{\sigma}_s^2 = \frac{1}{n_s - 1}\sum_{i\in s}(y_i - \bar{y}_s)^2$ and $Var(Y_i^* \mid s) \approx \hat{\sigma}_s^2$)

Proof of Lemma 1

We get

$$E(\hat{\theta}^* \mid s) = E\left(\sum_{i\in s_r} a_i(s)y_i + \sum_{i\in s - s_r} a_i Y_i^* \mid s\right)$$

$$= E_{s_r|s}E\left(\sum_{i\in s_r} a_i(s)y_i + \sum_{i\in s - s_r} a_i(s)Y_i^* \mid s, s_r\right)$$

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

First term :

$$E\left(\sum_{i\in s_r} a_i(s)y_i \mid s\right) = EE\left(\sum_{i\in s_r} a_i(s)y_i \mid s, n_r\right)$$

$$= EE\left(\sum_{i\in s} a_i(s)y_i R_i \mid s, n_r\right) = E\left(\sum_{i\in s} a_i(s)y_i P(R_i = 1 \mid s, n_r\right)$$

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

Second term:

$$E\left(\sum_{i\in s-s_r} a_i(s)\bar{y}_r \mid s\right) = EE\left(\sum_{i\in s-s_r} a_i(s)\bar{y}_r \mid s, n_r\right)$$

$$= EE\left(\frac{1}{n_r}\sum_{i\in s-s_r}\sum_{j\in s_r} a_i(s) y_j \mid s, n_r\right)$$

$$= EE\left(\frac{1}{n_r}\sum_{i\in s}\sum_{j\in s} a_i(s) y_j (1-R_i)R_j \mid s, n_r\right)$$

$$= E\left(\frac{1}{n_r}\sum_{i\in s}\sum_{\substack{j\in s\\ j\neq i}} a_i(s) y_j (E(R_j \mid s, n_r) - E(R_i R_j \mid s, n_r))\right)$$

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

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

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

This implies that

$$E(\hat{\theta}^* \mid s) = p_r\hat{\theta} + \frac{1-p_r}{n-1}(n^2\bar{a}(s)\bar{y}_s - \hat{\theta})$$

and (22) is equivalent to

$$p_r\hat{\theta} + \frac{1-p_r}{n-1}(n^2\bar{a}(s)\bar{y}_s - \hat{\theta}) = \hat{\theta} \Leftrightarrow \hat{\theta}\left(1 + \frac{1-p_r}{n-1} - p_r\right) = \frac{1-p_r}{n-1} n^2\bar{a}(s)\bar{y}_s$$

$$\Leftrightarrow \hat{\theta}\frac{n(1-p_r)}{n-1} = \frac{1-p_r}{n-1} n^2\bar{a}(s)\bar{y}_s \Leftrightarrow \hat{\theta} = n\bar{a}(s)\bar{y}_s = \bar{a}(s)\sum_{i\in s} y_i$$

and the result follows. □

Proof of Theorem

From Lemma 1, $\hat{\theta} = a(s)\sum_{i\in s} y_i = na(s)\bar{y}_s$ and $\hat{\theta}^* = a(s)\left(\sum_{i\in s_r} y_i + \sum_{i\in s-s_r} a_i y_i^*\right)$.

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

$$Var(\hat{\theta}^* \mid s, s_r) =^{(e)} [a(s)]^2 (n - n_r) \frac{n_r - 1}{n_r} \hat{\sigma}_r^2 .$$

Hence,

$$VarE(\hat{\theta}^* \mid s, s_r) = Var(na(s)\bar{y}_r) = EVar(na(s)\bar{y}_r \mid s) + VarE(na(s)\bar{y}_r \mid s)$$

$$= En^2[a(s)]^2 Var(\bar{y}_r \mid s) + Var\{na(s)E(\bar{y}_r \mid s)\}$$

$$= En^2[a(s)]^2 \{E_{n_r|s} Var(\bar{y}_r \mid s, n_r) + Var_{n_r|s} E(\bar{y}_r \mid s, n_r)\} + Var\{na(s)E_{n_r|s} E(\bar{y}_r \mid s, n_r)\}$$

$$=^{(b)} En^2[a(s)]^2 \{E_{n_r}(\hat{\sigma}_s^2(\frac{1}{n_r} - \frac{1}{n}) + Var_{n_r}(\bar{y}_s)\} + Var\ \{na(s)E_{n_r}\bar{y}_s\}$$

$$= En^2[a(s)]^2 \{(\hat{\sigma}_s^2 E(\frac{1}{n_r}) - \frac{1}{n}) + 0 + Var\ \{na(s)E_{n_r}\bar{y}_s\}$$

$$= n^2(E(\frac{1}{n_r}) - \frac{1}{n})E[a(s)]^2\hat{\sigma}_s^2 + Var\,\hat{\theta} .$$

Next,

$$EVar(\hat{\theta}^* \mid s, s_r) =^{(f)} E\{[a(s)]^2 (n - n_r) \frac{n_r - 1}{n_r} \hat{\sigma}_r^2\}$$

$$= EE\{[a(s)]^2 (n - n_r) \frac{n_r - 1}{n_r} \hat{\sigma}_r^2 \mid s, n_r\} \approx E\{[a(s)]^2 (n - n_r) E(\hat{\sigma}_r^2 \mid s, n_r)\}$$

$$=^{(b)} E[a(s)]^2 \hat{\sigma}_s^2 (n - n_r) = EE([a(s)]^2 \hat{\sigma}_s^2 (n - n_r) \mid s) = n(1 - p_r)E[a(s)]^2 \hat{\sigma}_s^2 .$$

We find now, from (21)

$$E(k) = \frac{VarE(\hat{\theta}^* \mid s, s_r) - Var(\hat{\theta})}{EVar(\hat{\theta}^* \mid s, s_r)}$$

$$= \frac{n^2(E(\frac{1}{n_r}) - \frac{1}{n})E[a(s)]^2\hat{\sigma}_s^2}{n(1 - p_r)E[a(s)]^2\hat{\sigma}_s^2} = \frac{n(E(1/n_r) - 1)}{1 - p_r} \approx \frac{(1/p_r) - 1}{1 - p_r} = \frac{1}{p_r} \quad . \qquad \square$$

Reference

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

Recent publications in the series Discussion Papers

331 M.W. Arneberg, J.K. Dagsvik and Z. Jia (2002): Labor Market Modeling Recognizing Latent Job Attributes and Opportunity Constraints. An Empirical Analysis of Labor Market Behavior of Eritrean Women

332 M. Greaker (2002): Eco-labels, Production Related Externalities and Trade

333 J. T. Lind (2002): Small continuous surveys and the Kalman filter

334 B. Halvorsen and T. Willumsen (2002): Willingness to Pay for Dental Fear Treatment. Is Supplying Fear Treatment Social Beneficial?

335 T. O. Thoresen (2002): Reduced Tax Progressivity in Norway in the Nineties. The Effect from Tax Changes

336 M. Søberg (2002): Price formation in monopolistic markets with endogenous diffusion of trading information: An experimental approach

337 A. Bruvoll og B.M. Larsen (2002): Greenhouse gas emissions in Norway. Do carbon taxes work?

338 B. Halvorsen and R. Nesbakken (2002): A conflict of interests in electricity taxation? A micro econometric analysis of household behaviour

339 R. Aaberge and A. Langørgen (2003): Measuring the Benefits from Public Services: The Effects of Local Government Spending on the Distribution of Income in Norway

340 H. C. Bjørnland and H. Hungnes (2003): The importance of interest rates for forecasting the exchange rate

341 A. Bruvoll, T.Fæhn and Birger Strøm (2003): Quantifying Central Hypotheses on Environmental Kuznets Curves for a Rich Economy: A Computable General Equilibrium Study

342 E. Biørn, T. Skjerpen and K.R. Wangen (2003): Parametric Aggregation of Random Coefficient Cobb-Douglas Production Functions: Evidence from Manufacturing Industries

343 B. Bye, B. Strøm and T. Åvitsland (2003): Welfare effects of VAT reforms: A general equilibrium analysis

344 J.K. Dagsvik and S. Strøm (2003): Analyzing Labor Supply Behavior with Latent Job Opportunity Sets and Institutional Choice Constraints

345 A. Raknerud, T. Skjerpen and A. Rygh Swensen (2003): A linear demand system within a Seemingly Unrelated Time Series Equation framework

346 B.M. Larsen and R.Nesbakken (2003): How to quantify household electricity end-use consumption

347 B. Halvorsen, B. M. Larsen and R. Nesbakken (2003): Possibility for hedging from price increases in residential energy demand

348 S. Johansen and A. R. Swensen (2003): More on Testing Exact Rational Expectations in Cointegrated Vector Autoregressive Models: Restricted Drift Terms

349 B. Holtsmark (2003): The Kyoto Protocol without USA and Australia - with the Russian Federation as a strategic permit seller

350 J. Larsson (2003): Testing the Multiproduct Hypothesis on Norwegian Aluminium Industry Plants

351 T. Bye (2003): On the Price and Volume Effects from Green Certificates in the Energy Market

352 E. Holmøy (2003): Aggregate Industry Behaviour in a Monopolistic Competition Model with Heterogeneous Firms

353 A. O. Ervik, E.Holmøy and T. Hægeland (2003): A Theory-Based Measure of the Output of the Education Sector

354 E. Halvorsen (2003): A Cohort Analysis of Household Saving in Norway

355 I. Aslaksen and T. Synnestvedt (2003): Corporate environmental protection under uncertainty

356 S. Glomsrød and W. Taoyuan (2003): Coal cleaning: A viable strategy for reduced carbon emissions and improved environment in China?

357 A. Bruvoll T. Bye, J. Larsson og K. Telle (2003): Technological changes in the pulp and paper industry and the role of uniform versus selective environmental policy.

358 J.K. Dagsvik, S. Strøm and Z. Jia (2003): A Stochastic Model for the Utility of Income.

359 M. Rege and K. Telle (2003): Indirect Social Sanctions from Monetarily Unaffected Strangers in a Public Good Game.

360 R. Aaberge (2003): Mean-Spread-Preserving Transformation.

361 E. Halvorsen (2003): Financial Deregulation and Household Saving. The Norwegian Experience Revisited

362 E. Røed Larsen (2003): Are Rich Countries Immune to the Resource Curse? Evidence from Norway's Management of Its Oil Riches

363 E. Røed Larsen and Dag Einar Sommervoll (2003): Rising Inequality of Housing? Evidence from Segmented Housing Price Indices

364 R. Bjørnstad and T. Skjerpen (2003): Technology, Trade and Inequality

365 A. Raknerud, D. Rønningen and T. Skjerpen (2003): A method for improved capital measurement by combining accounts and firm investment data

366 B.J. Holtsmark and K.H. Alfsen (2004): PPP-correction of the IPCC emission scenarios - does it matter?

367 R. Aaberge, U. Colombino, E. Holmøy, B. Strøm and T. Wennemo (2004): Population ageing and fiscal sustainability: An integrated micro-macro analysis of required tax changes

368 E. Røed Larsen (2004): Does the CPI Mirror Costs.of.Living? Engel's Law Suggests Not in Norway

369 T. Skjerpen (2004): The dynamic factor model revisited: the identification problem remains

370 J.K. Dagsvik and A.L. Mathiassen (2004): Agricultural Production with Uncertain Water Supply

371 M. Greaker (2004): Industrial Competitiveness and Diffusion of New Pollution Abatement Technology – a new look at the Porter-hypothesis

372 G. Børnes Ringlund, K.E. Rosendahl and T. Skjerpen (2004): Does oilrig activity react to oil price changes? An empirical investigation

373 G. Liu (2004) Estimating Energy Demand Elasticities for OECD Countries. A Dynamic Panel Data Approach

374 K. Telle and J. Larsson (2004): Do environmental regulations hamper productivity growth? How

accounting for improvements of firms' environmental performance can change the conclusion

375 K.R. Wangen (2004): Some Fundamental Problems in Becker, Grossman and Murphy's Implementation of Rational Addiction Theory

376 B.J. Holtsmark and K.H. Alfsen (2004): Implementation of the Kyoto Protocol without Russian participation

377 E. Røed Larsen (2004): Escaping the Resource Curse and the Dutch Disease? When and Why Norway Caught up with and Forged ahead of Its Neughbors

378 L. Andreassen (2004): Mortality, fertility and old age care in a two-sex growth model

379 E. Lund Sagen and F. R. Aune (2004): The Future European Natural Gas Market - are lower gas prices attainable?

380 A. Langørgen and D. Rønningen (2004): Local government preferences, individual needs, and the allocation of social assistance

381 K. Telle (2004): Effects of inspections on plants' regulatory and environmental performance - evidence from Norwegian manufacturing industries

382 T. A. Galloway (2004): To What Extent Is a Transition into Employment Associated with an Exit from Poverty

383 J. F. Bjørnstad and E.Ytterstad (2004): Two-Stage Sampling from a Prediction Point of View

384 A. Bruvoll and T. Fæhn (2004): Transboundary environmental policy effects: Markets and emission leakages

385 P.V. Hansen and L. Lindholt (2004): The market power of OPEC 1973-2001

386 N. Keilman and D. Q. Pham (2004): Empirical errors and predicted errors in fertility, mortality and migration forecasts in the European Economic Area

387 G. H. Bjertnæs and T. Fæhn (2004): Energy Taxation in a Small, Open Economy: Efficiency Gains under Political Restraints

388 J.K. Dagsvik and S. Strøm (2004): Sectoral Labor Supply, Choice Restrictions and Functional Form

389 B. Halvorsen (2004): Effects of norms, warm-glow and time use on household recycling

390 I. Aslaksen and T. Synnestvedt (2004): Are the Dixit-Pindyck and the Arrow-Fisher-Henry-Hanemann Option Values Equivalent?

391 G. H. Bjønnes, D. Rime and H. O.Aa. Solheim (2004): Liquidity provision in the overnight foreign exchange market

392 T. Åvitsland and J. Aasness (2004): Combining CGE and microsimulation models: Effects on equality of VAT reforms

393 M. Greaker and Eirik. Sagen (2004): Explaining experience curves for LNG liquefaction costs: Competition matter more than learning

394 K. Telle, I. Aslaksen and T. Synnestvedt (2004): "It pays to be green" - a premature conclusion?

395 T. Harding, H. O. Aa. Solheim and A. Benedictow (2004). House ownership and taxes

396 E. Holmøy and B. Strøm (2004): The Social Cost of Government Spending in an Economy with Large Tax Distortions: A CGE Decomposition for Norway

397 T. Hægeland, O. Raaum and K.G. Salvanes (2004): Pupil achievement, school resources and family background

398 I. Aslaksen, B. Natvig and I. Nordal (2004): Environmental risk and the precautionary principle: "Late lessons from early warnings" applied to genetically modified plants

399 J. Møen (2004): When subsidized R&D-firms fail, do they still stimulate growth? Tracing knowledge by following employees across firms

400 B. Halvorsen and Runa Nesbakken (2004): Accounting for differences in choice opportunities in analyses of energy expenditure data

401 T.J. Klette and A. Raknerud (2004): Heterogeneity, productivity and selection: An empirical study of Norwegian manufacturing firms

402 R. Aaberge (2005): Asymptotic Distribution Theory of Empirical Rank-dependent Measures of Inequality

403 F.R. Aune, S. Kverndokk, L. Lindholt and K.E. Rosendahl (2005): Profitability of different instruments in international climate policies

404 Z. Jia (2005): Labor Supply of Retiring Couples and Heterogeneity in Household Decision-Making Structure

405 Z. Jia (2005): Retirement Behavior of Working Couples in Norway. A Dynamic Programming Approch

406 Z. Jia (2005): Spousal Influence on Early Retirement Behavior

407 P. Frenger (2005): The elasticity of substitution of superlative price indices

408 M. Mogstad, A. Langørgen and R. Aaberge (2005): Region-specific versus Country-specific Poverty Lines in Analysis of Poverty

409 J.K. Dagsvik (2005) Choice under Uncertainty and Bounded Rationality

410 T. Fæhn, A.G. Gómez-Plana and S. Kverndokk (2005): Can a carbon permit system reduce Spanish unemployment?

411 J. Larsson and K. Telle (2005): Consequences of the IPPC-directive's BAT requirements for abatement costs and emissions

412 R. Aaberge, S. Bjerve and K. Doksum (2005): Modeling Concentration and Dispersion in Multiple Regression

413 E. Holmøy and K.M. Heide (2005): Is Norway immune to Dutch Disease? CGE Estimates of Sustainable Wage Growth and De-industrialisation

414 K.R. Wangen (2005): An Expenditure Based Estimate of Britain's Black Economy Revisited

415 A. Mathiassen (2005): A Statistical Model for Simple, Fast and Reliable Measurement of Poverty

416 F.R. Aune, S. Glomsrød, L. Lindholt and K.E. Rosendahl: Are high oil prices profitable for OPEC in the long run?

417 D. Fredriksen, K.M. Heide, E. Holmøy and I.F. Solli (2005): Macroeconomic effects of proposed pension reforms in Norway

418 D. Fredriksen and N.M. Stølen (2005): Effects of demographic development, labour supply and pension reforms on the future pension burden

419 A. Alstadsæter, A-S. Kolm and B. Larsen (2005): Tax Effects on Unemployment and the Choice of Educational Type

420 E. Biørn (2005): Constructing Panel Data Estimators by Aggregation: A General Moment Estimator and a Suggested Synthesis

421 J. Bjørnstad (2005): Non-Bayesian Multiple Imputation