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 \( y = (y_1, \ldots, y_n) \) denoting the full sample data, values of random variable \( Y_1, \ldots, Y_n \). The objective is to estimate some parameter \( \theta \). 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 \( \hat{V}(y) \). 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 \( \hat{\theta}^* = \hat{\theta}(y^*) \) and \( \hat{V}^* = \hat{V}(y^*) \).

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

\[
\bar{\theta}^* = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i^* / m.
\]

The within-imputation variance is defined as

\[
\bar{V}^* = \frac{1}{m} \sum_{i=1}^{m} \hat{V}_i^* / m
\]

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

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

That is, we need to determine \( k \) such that

\[
E(W) = Var(\bar{\theta}^*).
\]

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,

\[
E(W) = E\{E(W | y_{obs})\}
\]
\[ \text{Var}(\theta^*) = E\{\text{Var}(\theta^* | Y_{\text{obs}})\} + \text{Var}\{E(\theta^* | Y_{\text{obs}})\}. \]

Typically,

\[ E(\tilde{V}^*) \approx \text{Var}(\hat{\theta}) \quad (3) \]

and

\[ E(B^* | y_{\text{obs}}) = \text{Var}(\tilde{\theta}^* | y_{\text{obs}}). \]

Hence, approximately

\[ E(W) = \text{Var}(\hat{\theta}) + (E(k) + \frac{1}{m})E\text{Var}(\tilde{\theta}^* | Y_{\text{obs}}). \quad (4) \]

Moreover,

\[ \text{Var}(\tilde{\theta}^* | y_{\text{obs}}) = \text{Var}(\tilde{\theta}^* | y_{\text{obs}}) / m \]

and

\[ E(\tilde{\theta}^* | y_{\text{obs}}) = E(\tilde{\theta}^* | y_{\text{obs}}). \]

This implies that

\[ \text{Var}(\tilde{\theta}^*) = \frac{1}{m} E\{\text{Var}(\tilde{\theta}^* | Y_{\text{obs}})\} + \text{Var}\{E(\tilde{\theta}^* | Y_{\text{obs}})\}. \]

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

\[ \text{Var}(\hat{\theta}) + E(k)E\text{Var}(\tilde{\theta}^* | Y_{\text{obs}}) = \text{Var}\{E(\tilde{\theta}^* | Y_{\text{obs}})\}, \]

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

\[ E(k) = \frac{\text{Var}E(\tilde{\theta}^* | Y_{\text{obs}}) - \text{Var}(\hat{\theta})}{E\text{Var}(\tilde{\theta}^* | Y_{\text{obs}})}. \quad (5) \]

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 \( \mu \) 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 \( R_1, \ldots, R_N \) 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 \( y_{obs} \), and the estimate is the sample mean. Let \( \bar{y}_r \) be the observed sample mean and \( \sigma_r^2 = \frac{1}{n_r} \sum_{i \in s_r} (y_i - \bar{y}_r)^2 \) the observed sample variance. Then \( \bar{Y}^* \) is the imputation-based sample mean for the completed sample, and the combined estimator is given by

\[
\bar{Y}^* = \frac{\sum_{i=1}^{m} \bar{Y}_i}{m}.
\]

Let \( \bar{Y}_s \) denote the sample mean based on a full sample. Then,

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

being the population variance. We have further that

\[
E(\bar{Y}^* | y_{obs}) = \bar{y}_r \quad \text{and} \quad Var(\bar{Y}^* | y_{obs}) = \frac{n-n_r}{n^2} \cdot \frac{n_r-1}{n_r} \sigma_r^2.
\]

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

In this case,

\[
\hat{\sigma}_r^2 = \frac{1}{n-1} \left( \sum_{i \in s_r} (y_i - \bar{y}_r)^2 + \sum_{i \in s_{-r}} (y_i^* - \bar{y}_r)^2 \right).
\]

It can be shown that

\[
E(\hat{\sigma}_r^2 | y_{obs}) = \hat{\sigma}_r^2 (1 - \frac{1}{n}) \left( 1 + \frac{n}{n(n-1)} \right) \approx \hat{\sigma}_r^2
\]

and (3) holds. We find, from (5),
\[ E(k) = \frac{\text{Var}(\bar{Y}_r) - \sigma^2 \left( \frac{1}{n} - \frac{1}{n_r} \right)}{E\left( \frac{n-n_r}{n} \cdot \frac{n_i-1}{n_i} \right) E(\sigma_r^2 | n_r)} \]
\[ = \frac{\sigma^2 \left( E\left( \frac{1}{n} \right) - \frac{1}{n} - \frac{1}{n_r} \right)}{E\left( \frac{n-n_r}{n} \cdot \frac{n_r-1}{n_r} \right) \sigma^2} \]
\[ \approx \frac{(1-p_r)/p_r}{1-p_r} = \frac{1}{p_r} \]
which is satisfied approximately 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:
\[ Y_i = \beta x_i + \epsilon_i \], with \( \text{Var}(\epsilon_i) = \sigma^2 x_i \); \( i = 1, \ldots, n \).
It is assumed that all \( x_i \)’s are known, also in the nonresponse sample. The full data estimator of \( \beta \) is given by
\[ \hat{\beta} = \frac{\sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i} . \]
The unbiased estimator of \( \sigma^2 \) is given by
\[ \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \frac{1}{x_i} (y_i - \hat{\beta} x_i)^2 . \]
We shall consider residual regression imputation:
Let \( \hat{\beta} \) be the \( \hat{\beta} \) - estimate based on observed sample \( s_r \). Define the standardized residuals
\[ \epsilon_i = (y_i - \hat{\beta} x_i) / \sqrt{x_i} \], for \( i \in s_r \).
For \( i \in s - s_r \) : Draw the value of \( \epsilon_i^* \) at random from the set of observed residuals \( \epsilon_i, i \in s_r \), and the imputed \( y \)-value is given by
\[ y_i^* = \hat{\beta} x_i + \epsilon_i^* \sqrt{x_i} . \]
Let \( X = \sum_{i=1}^{n} x_i, X_r = \sum_{i \in r} x_i \) and \( X_{nr} = \sum_{i \in \overline{r}} x_i = X - X_r \). 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:

\[
f_X = \frac{X_{nr}}{X}.
\]

We now have

\[
\hat{\beta}^* = \left( \sum_s y_i + \sum_{s \neq s'} y_i^* \right) / X
\]

\[
\hat{\alpha}^2 = \frac{1}{n - 1} \left( \sum_{s} \frac{1}{x_i} (y_i - \hat{\beta}^* x_i)^2 + \sum_{s \neq s'} \frac{1}{x_i} (y_i^* - \hat{\beta}^* x_i)^2 \right).
\]

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

\[
\text{Var}(\hat{\beta}) = \sigma^2 / X.
\]

Consider (3) which is equivalent to

\[
\text{Var}(\hat{\beta}^* | y_{\text{obs}}) \approx \sigma^2.
\]

Let \( \hat{\beta}_{nr} = \sum_{i \in \overline{r}} y_i^* / X_{nr} \), and \( \hat{\alpha}_{nr}^2 = \frac{1}{n_{nr} - 1} \sum_{i \in \overline{r}} \frac{1}{x_i} (y_i^* - \hat{\beta}_{nr} x_i)^2 \). Here, \( n_{nr} = n - n_r \). Then, after some algebra, one can express \( \hat{\alpha}^2 \) in the following way:

\[
\hat{\alpha}^2 = \frac{1}{n - 1} \left( (n_r - 1) \hat{\alpha}^2 + (n_{nr} - 1) \hat{\alpha}_{nr}^2 + \frac{X_{nr} X}{X} (\hat{\beta} - \hat{\beta}_{nr})^2 \right).
\]

In this case,

\[
\text{E}(Y_i^* | y_{\text{obs}}) = \hat{\beta}_r x_i + \overline{\epsilon} \sqrt{x_i}, \text{ where } \overline{\epsilon} = \sum_{i} \frac{e_i}{n_r},
\]

\[
\text{Var}(Y_i^* | y_{\text{obs}}) = x_i s_c^2, \text{ where } s_c^2 = \frac{1}{n_r - 1} \sum_{i \in \overline{r}} (e_i - \overline{\epsilon})^2.
\]

Using this, it can be shown that

\[
\text{E}(\hat{\alpha}^2) = \sigma^2 \left( 1 - \frac{c_1}{n - 1} - \frac{4c_2}{(n - 1)n_r} - c_3 f \frac{n - 1}{n \cdot n_r} \right)
\]

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

Hence, \( \text{E}(\hat{\alpha}^2) \approx \sigma^2 \) and (3) follows, at least for moderate and large \( n_r \).
Next, we look at $\text{Var}(\hat{\beta}^* \mid y_{\text{obs}})$ and $E(\hat{\beta}^* \mid y_{\text{obs}})$:

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

$$E(\hat{\beta}_{nr} \mid y_{\text{obs}}) = \hat{\beta}_r + \frac{\bar{e}}{X_{nr} s^2} \sum \sqrt{x_i}$$

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

This gives us

$$E(\hat{\beta}^* \mid y_{\text{obs}}) = \hat{\beta}_r + \frac{\bar{e}}{X s^2} \sum \sqrt{x_i}$$

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

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

$$\text{Var}(\hat{\beta}^* \mid y_{\text{obs}}) = \text{Var}(\hat{\beta}_r) + \left( \frac{\sum_{i=s}^n \sqrt{x_i}^2}{X^2} \right) \text{Var}(\bar{e}) + 2 \frac{\sum_{s=s_r}^n \sqrt{x_i}}{X} \text{Cov}(\hat{\beta}_r, \bar{e}).$$

Using Cauchy-Schwarz inequality,

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

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

$$\left( \sum_{i=1}^n \sqrt{x_i} \right)^2 \leq n X. \quad (6)$$

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

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

Moreover, from (6),

$$\frac{\sum_{s=s_r}^n \sqrt{x_i}^2}{X^2} = d_2 n_{nr} X_{nr} / X^2, \quad 0 \leq d_2 \leq 1.$$
Hence,
\[
\text{Var}E(\hat{\beta}^* | y_{\text{obs}}) = \frac{\sigma^2}{X_r} + \frac{(1-d_1)d_2n_rX_{nr}}{X_r^2} \frac{\sigma^2}{n_r}.
\]

Next we find that
\[
E(s_{\text{err}}^2) = \sigma^2(1 - \frac{1}{n_r}) - \text{Var}(\bar{v}) = \frac{\sigma^2}{n_r}(n_r + d_1 - 2)
\]
which gives us
\[
E\text{Var}(\hat{\beta}^* | y_{\text{obs}}) = \frac{X_{nr}}{X_r^2} \frac{\sigma^2}{n_r}(n_r + d_1 - 2).
\]

From (5),
\[
k = \frac{\sigma^2}{X_r} + \frac{\sigma^2}{n_r} \frac{(1-d_1)d_2n_rX_{nr}}{X_r^2} - \frac{\sigma^2}{X_r} \frac{n_r}{X_r^2}(n_r + d_1 - 2)
\]
\[
= \frac{n_rX^2 - n_rX \cdot X_r + (1-d_1)d_2n_rX_{nr}X_r}{X_rX_{nr}(n_r + d_1 - 2)}
\]
\[
\approx \frac{X}{X_r} + (1-d_1)d_2 \frac{n_{nr}}{n_r}.
\]

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.
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 $y_h$ be the planned full data from sub sample $s_h$ of size $n_h$. It is assumed that $y_1, \ldots, y_H$ are independent.

The observed part of $y_h$ is denoted by $y_{h,\text{obs}}$ 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 $y_h$.

$Var(\hat{\theta}) = \sum_{h=1}^H Var(\hat{\theta}_h)$ 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 $y_h$. For $i \in s_h - s_{hr}$ we impute by some method $y^*_i$ based on $y_{h,\text{obs}}$ and let $y_h^*$ denote the complete data $(y_{h,\text{obs}}, y^*_i, i \in s_h - s_{hr})$. Based on $y_h^*$, we have $\hat{\theta}_h^* = \hat{\theta}_h(y_h^*)$ and $\hat{V}_h^* = \hat{V}_h(y_h^*)$. 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 $\bar{\theta}_i = \sum_{h=1}^H \hat{\theta}_{h,i}$ and $\bar{V}_i = \sum_{h=1}^H \hat{V}_{h,i}$, for $i = 1,\ldots,m$. The combined estimate for stratum $h$ is given by

$$\bar{\theta}_h = \frac{\sum_{i=1}^m \hat{\theta}_{h,i}}{m}.$$ 

The within-imputation variance for stratum $h$ is

$$\bar{V}_h^* = \frac{\sum_{i=1}^m V_{h,i}^*}{m}$$

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

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

where

$$\bar{V}^* = \sum_{i=1}^{m} \hat{V}_i^* / m = \sum_{h=1}^{H} \bar{V}_h^*.$$ 

Provided (3) holds for each stratum $$h$$,

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

we have from (5) that $$k_h$$ must satisfy

$$E(k_h) = \frac{VarE(\hat{\theta}_h \mid Y_{h,obs}) - Var(\hat{\theta}_h)}{EV(\hat{\theta}_h \mid Y_{h,obs}).}$$

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 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

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

Here, $$B^* = \frac{1}{m-1} \sum_{i=1}^{m} (\hat{\theta}_i^* - \bar{\theta}^*)^2 = \frac{1}{m-1} \sum_{i=1}^{m} \left( \sum_{h=1}^{H} (\hat{\theta}_{ih}^* - \bar{\theta}_h^*) \right)^2.$$ We note that
\[
E(B^* \mid y_{obs}) = E(\sum_{h=1}^H B_h^* \mid y_{obs}).
\]

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):
\[
k = \frac{\sum_{h=1}^H k_h E(B_h^* \mid y_{obs})}{E(B^* \mid y_{obs})}.
\]
\[
= \frac{\sum_{h=1}^H k_h Var(\hat{\theta}_h^* \mid y_{obs})}{Var(\hat{\theta}^* \mid y_{obs})} = \sum_{h=1}^H k_h \cdot \frac{Var(\hat{\theta}_h^* \mid y_{obs})}{Var(\hat{\theta}^* \mid y_{obs})},
\]

(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 \(N_h\), \(h = 1, \ldots, H\). The aim is to estimate the population average \(\mu\) 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 \(h\), \(R_{h,1}, \ldots, R_{h,N_h}\) 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 \( y_{h,\text{obs}} \).

The estimator based on the full sample data is the usual stratified weighted average

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

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

\[
\text{Var}(\bar{Y}_{\text{strat}}) = \sum_{h=1}^{H} v_h^2 \sigma_h^2 \left( \frac{1}{n_h} - \frac{1}{N_h} \right) , \quad \text{with} \quad \sigma_h^2 = \frac{1}{N_h-1} \sum_{i \in s_h} (y_i - \mu_h)^2
\]

being the population variance in stratum \( h \). Here \( U_h \) is stratum population \( h \) and \( \mu_h \) is the average in \( U_h \).

Let \( \bar{y}_h \) 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

\[
\bar{Y}_{\text{strat}}^* = \frac{1}{N} \sum_{h=1}^{H} N_k \bar{y}_h^* 
\]

where

\[
\bar{y}_h^* = \frac{1}{n_h} \left( \sum_{i \in s_{hr}} y_i + \sum_{i \in s_{hr}} y_i^* \right) = \frac{1}{n_h} (n_{hr} \bar{y}_{hr} + \sum_{i \in s_{hr}} y_i^* ) .
\]

Let the \( m \) imputation replicates of \( \bar{Y}_{\text{strat}}^* \) be denoted by \( \bar{Y}_{\text{strat},i}^* \) for \( i = 1, \ldots, m \). The combined estimator is given by

\[
\bar{Y}_{\text{strat}}^* = \frac{1}{m} \sum_{i=1}^{m} \bar{Y}_{\text{strat},i}^* .
\]

### 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 \( \bar{Y}_{\text{strat}}^* \) becomes, from (7),

\[
W_{\text{sep}} = \bar{V}^* + \sum_{h=1}^{H} \left( \frac{1}{1-f_h} + \frac{1}{m} \right) B_h^* .
\]
Here, $\hat{F}^* = \sum_{h=1}^{H} F_h^*$ and $F_h^*$ is the average of the $m$ values of the imputation based variance estimate

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

where

$$\tilde{\sigma}_{he}^2 = \frac{1}{n_h - 1} \left( \sum_{i \in s_h} (y_i - \bar{y}_h^*)^2 + \sum_{h \neq s_h} (y_i^* - \bar{y}_h^*)^2 \right).$$

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

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

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

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

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

This gives the following results:

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

Hence we can determine $k$ as

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

If the stratum sizes $N_h$ are large then we can let $\hat{V}(v_h F^*_h) = v_h^2 \hat{\sigma}_{hr}^2 / n_h$. Let also

$$b_h = f_h \hat{V}(v_h F^*_h) / \sum_{h=1}^{H} f_k \hat{V}(v_k F^*_k).$$ 

Then

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

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 \({\hat{Y}}_h\) are independent which implies that

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

It follows that

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

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

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

\[
= \frac{\sum_{h=1}^{H} v_h^2 \sigma^2_h [E(\frac{1}{n_h}) - \frac{1}{N}] - \sum_{h=1}^{H} v_h^2 \sigma^2_h (\frac{1}{n_h} - \frac{1}{N})}{\sum_{h=1}^{H} v_h^2 \cdot \sigma^2_h \frac{E(f_h)}{n_h}}
\]

\[
\approx \frac{\sum_{h=1}^{H} v_h^2 \sigma^2_h \frac{1 - p_{hr}}{n_h} \cdot \frac{1}{p_{hr}}}{\sum_{h=1}^{H} v_h^2 \sigma^2_h \frac{E(f_h)}{n_h}} - \frac{\sum_{h=1}^{H} v_h^2 \sigma^2_h E(f_h) \frac{1 - f_h}{n_{hr}}}{\sum_{h=1}^{H} v_h^2 \sigma^2_h E(f_h)(1 - f_h)}.
\]

\[\text{(13)}\]

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

\[
\frac{1}{k} = \frac{\sum_{h=1}^{H} (1 - f_h) f_h \hat{V}(v_h \overline{Y}_{hr})}{\sum_{h=1}^{H} f_h \hat{V}(v_h \overline{Y}_{hr})} = \sum_{h=1}^{H} a_h (1 - f_h).
\]

\[\text{(14)}\]
where the weights \( a_h = f_h \hat{V}(v_h Y_{hr}) / \sum_{h=1}^{H} f_h \hat{V}(v_h Y_{hr}) \). 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 \( v_h Y_{hr} \) is large. We note that the estimate of the total based on the response sample is given by

\[
\bar{Y}_{strat,r} = \sum_{h=1}^{H} v_h \bar{Y}_{hr}.
\]

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

\[
E(k) \approx \frac{\sum_{h=1}^{H} \text{Var}(v_h Y_{hr}) E(f_h)}{\sum_{h=1}^{H} \text{Var}(v_h Y_{hr}) E(f_h)}.
\]

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:

- \( Y_1, \ldots, Y_n \) are independent 0/1 -variables
- Explanatory 0/1-variable \( x \) with fixed known values \( x_1, \ldots, x_n \)
- Class probabilities: \( \pi_0 = P(Y_j = 1 \mid x_i = 1) \) and \( \pi_0 = P(Y_j = 1 \mid x_i = 0) \)
- Response variables: \( R_1, \ldots, R_n \) with MAR (missing at random) model:
  \[
P(R_i = 1 \mid x_i = 1) = p_{1r} \quad \text{and} \quad P(R_i = 1 \mid x_i = 0) = p_{0r}
\]

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:

\[
\alpha = \log \frac{\pi_0}{1-\pi_0} = \frac{1}{1+e^{-\alpha}}
\]

\[
\beta = \log \frac{\pi_0/(1-\pi_0)}{\pi_0/(1-\pi_0)} = \log(\text{odds ratio}), \quad \text{and} \quad \pi_0 = \frac{1}{1+e^{-\beta}}.
\]

The aim is to estimate \( \beta \). 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 \( n_i = \sum_{i=1}^{n} x_i = X \) and
The response samples in the strata are \( s_r = \{ i \in s : R_i = 1 \} \) and \( s_{r_0} = \{ i \in s_0 : R_i = 1 \} \) with total response sample being \( s_r \) of size \( n_r \). Let also \( n_{r_r} = |s_r| \) and \( n_{r_0} = |s_{r_0}| \). We see that 

\[ n_{r_r} = \sum_{x = 0}^{1} x = X_r \quad \text{and} \quad n_{r_0} = n_r - X_r. \]

The data from \( s_r \) can be represented as follows where \( n_{ijr} \) denotes the number of observations with \( x = i \) and \( y = j \):

<table>
<thead>
<tr>
<th>( x \backslash y )</th>
<th>( y = 0 )</th>
<th>( y = 1 )</th>
<th>Totals</th>
<th>Nonresponse</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = 0 )</td>
<td>( n_{00r} )</td>
<td>( n_{01r} )</td>
<td>( n_{0r} )</td>
<td>( n_0 - n_{0r} )</td>
</tr>
<tr>
<td>( x = 1 )</td>
<td>( n_{10r} )</td>
<td>( n_{11r} )</td>
<td>( n_{1r} )</td>
<td>( n_1 - n_{1r} )</td>
</tr>
</tbody>
</table>

We then have the (maximum likelihood) estimates MLE)

\[ \hat{\pi}_{r_r} = n_{1r} / n_r \quad \text{and} \quad \hat{\pi}_{r_0r} = n_{0r} / n_{0r}. \]

Hence, MLE of \( \beta \) equals

\[ \hat{\beta}_r = \log \left( \frac{\hat{\pi}_{r_r}}{\hat{\pi}_{r_0r}} \right) = \log \frac{n_{1r}n_{00r}}{n_{10r}n_{01r}}. \]

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

\[ \hat{\beta} = \log \left( \frac{\hat{\pi}_1}{\hat{\pi}_0} \right) = \log \frac{n_{10}n_{00}}{n_{01}n_{00}} \]

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

\[ \hat{\beta} = \log \left( \frac{\hat{\pi}_1}{\hat{\pi}_0} \right) = \log \frac{\hat{\pi}_1}{1 - \hat{\pi}_1} - \log \frac{\hat{\pi}_0}{1 - \hat{\pi}_0} = \hat{\beta}_1 - \hat{\beta}_0, \]

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 \( N(\beta, \sigma^2) \) where

\[ \sigma^2 = \frac{1}{n_1 \pi_1 (1 - \pi_1)} + \frac{1}{n_0 \pi_0 (1 - \pi_0)}. \]

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

\[ \hat{\nu}(\hat{\beta}) = \frac{1}{n_1 \hat{\pi}_1 (1 - \hat{\pi}_1)} + \frac{1}{n_0 \hat{\pi}_0 (1 - \hat{\pi}_0)} \]

\[ = \frac{n_1}{n_1 n_{10}} + \frac{n_0}{n_0 n_{00}} \]
It follows that \( \hat{V}(\hat{\beta}) = \hat{V}_1 + \hat{V}_0 \), where \( \hat{V}_1 = \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \) and \( \hat{V}_0 = \left( \frac{1}{n_0} + \frac{1}{n_0} \right) \) 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 \( x = 1 \):

\[
y^* = 1 \text{ with probability } \hat{\pi}_{1r} = \frac{n_{11r}}{n_{1r}} \text{ and } y^* = 0 \text{ with probability } 1 - \hat{\pi}_{1r} = \frac{n_{01r}}{n_{1r}}.
\]

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 \( y_{i,obs} = (y_i : i \in s_{ir}) \) 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 \( y = j \):

<table>
<thead>
<tr>
<th>( x \backslash y )</th>
<th>( y = 0 )</th>
<th>( y = 1 )</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = 0 )</td>
<td>( n_{00}^* )</td>
<td>( n_{01}^* )</td>
<td>( n_0 - n_{0r} )</td>
</tr>
<tr>
<td>( x = 1 )</td>
<td>( n_{10}^* )</td>
<td>( n_{11}^* )</td>
<td>( n_1 - n_{1r} )</td>
</tr>
</tbody>
</table>

The imputation based estimate of \( \pi_1 \) is given by \( \hat{\pi}_1^* = \frac{n_{11r} + n_{11}^*}{n_1 - n_{1r} - n_{11}^*} \) such that the imputation based estimate \( \hat{\beta}_1^* \) becomes

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

Similarly, the imputation based estimates for \( \beta_0 \) and \( \beta \) 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 \( \hat{\beta}_1^* \), \( \hat{\beta}_0^* \), \( \hat{\beta}^* \), for \( i = 1, \ldots, m \). The combined estimate is given by

\[
\overline{\beta}^* = \frac{\sum_{i=1}^{m} \hat{\beta}_i^*}{m} = \frac{\sum_{i=1}^{m} \hat{\beta}_1^*}{m} - \frac{\sum_{i=1}^{m} \hat{\beta}_0^*}{m} = \overline{\beta}_1^* - \overline{\beta}_0^*.
\]

The imputed variance estimate \( \hat{V}^* \) for \( \hat{\beta} \) is given by

\[
\hat{V}^* = \frac{1}{n_{11r} + n_{11}^*} + \frac{1}{n_{10r} + n_{10}^*} + \frac{1}{n_{01r} + n_{01}^*} + \frac{1}{n_{00r} + n_{00}^*}.
\] (15)
We see that \( E(\hat{V}^* | y_{obs}) = \frac{1}{n_1 \hat{\pi}_{1r}(1 - \hat{\pi}_{1r})} + \frac{1}{n_0 \hat{\pi}_{0r}(1 - \hat{\pi}_{0r})} \) 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 \( E(\hat{\beta}_1^* | y_{1,obs}) \) and \( Var(\hat{\beta}_1^* | y_{1,obs}) \). We have that conditional on \( y_{1,obs}, n_1^* \) is binomially distributed \( (n_1 - n_r, \hat{\pi}_{1r}) \). Hence,

\[
E(n_1^* | y_{1,obs}) = (n_1 - n_r) \hat{\pi}_{1r} \quad \text{and} \quad Var(n_1^* | y_{1,obs}) = (n_1 - n_r) \hat{\pi}_{1r}(1 - \hat{\pi}_{1r}).
\]

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

\[
T = \log \frac{a + Z}{b - Z}, \quad \text{where} \ Z \ \text{is binomial} \ (n, \pi) \ \text{and} \ a \ \text{and} \ b \ \text{are constants}.
\]

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

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

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

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

and

\[
Var(\hat{\beta}_1^* | y_{1,obs}) \approx \left( \frac{n_1}{n_1 \hat{\pi}_{1r} n_1(1 - \hat{\pi}_{1r})} \right)^2 (n_1 - n_r) \hat{\pi}_{1r}(1 - \hat{\pi}_{1r}).
\]

Let \( f_i \) be the nonresponse rate in stratum \( s_i \); \( f_i = (n_i - n_{ir})/n_1 \). We see that

\[
Var(\hat{\beta}_1^* | y_{1,obs}) \approx \frac{f_i n_1}{n_i^2} \frac{1}{\hat{\pi}_{1r}(1 - \hat{\pi}_{1r})} = f_i (1 - f_i) \cdot \frac{1}{n_{ir} \hat{\pi}_{ir}(1 - \hat{\pi}_{ir})} = f_i (1 - f_i) \hat{V} (\hat{\beta}_{1r}).
\]
From (9), we find approximately:

\[
E(k_i) = \frac{\text{Var}(\hat{\beta}_{ir}) - \text{Var}(\hat{\beta}_i)}{E\{f_i(1 - f_i)\hat{V}(\hat{\beta}_{ir})\}}
\]

\[
= \frac{E\text{Var}(\hat{\beta}_{ir} | n_{ir}) - \text{Var}(\hat{\beta}_i)}{E\{f_i(1 - f_i)E[\hat{V}(\hat{\beta}_{ir}) | n_{ir}]\}}
\]

\[
\approx \frac{\frac{1}{n_i}E(\frac{1}{m_\pi} - \frac{1}{n})}{Ef_i(1 - f_i)\frac{1}{n_i}\pi(1 - \pi)}
\]

\[
\approx \frac{\frac{1}{n_i} - \frac{1}{m_\pi}}{\frac{1}{Ef_i}} = \frac{(1 - p_{ir})/p_{ir}}{1 - p_{ir}} = \frac{1}{p_{ir}}
\]

which is satisfied approximately by letting

\[
k_i = \frac{1}{1 - f_i}.
\]

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 \(\hat{\beta}_i^*\) is given by \(B_i^* = \frac{1}{m_\pi} \sum_{x=0}^{m_\pi} (\hat{\beta}_{1x}^* - \bar{\beta}_i^*)^2\) 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 \(\beta\) is given by, from (7),

\[
W_{sep} = \bar{V}^* + \sum_{x=0}^{m_\pi} \left(\frac{1}{1 - f_x} + \frac{1}{m}\right)B_x^*
\]

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

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

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

\[
\text{Var}(\hat{\beta}_1^* | y_{i,obs}) = f_i(1 - f_i)\hat{V}(\hat{\beta}_{ir})
\]

\[
\text{Var}(\hat{\beta}_0^* | y_{0,obs}) = f_0(1 - f_0)\hat{V}(\hat{\beta}_{0r}).
\]
It follows from (11) that
\[ k = \frac{1}{1-f_1} \cdot \frac{f_1(1-f_1)\hat{V}^2(\hat{\beta}_{1r})}{\sum_{r=0}^{1} f_r(1-f_r)\hat{V}^2(\hat{\beta}_{r})} + \frac{1}{1-f_0} \cdot \frac{f_0(1-f_0)\hat{V}^2(\hat{\beta}_{0r})}{\sum_{r=0}^{1} f_r(1-f_r)\hat{V}^2(\hat{\beta}_{r})}. \] (17)

Now, \( \text{Var}(\hat{\beta}_1) \approx \frac{n_{1r}}{n_r} \text{Var}(\hat{\beta}_{1r} | n_{1r}) = (1-f_1)\text{Var}(\hat{\beta}_{1r} | n_{1r}) \). Similarly, \( \text{Var}(\hat{\beta}_0) \approx (1-f_0)\text{Var}(\hat{\beta}_{0r} | n_{0r}) \). 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
\[ k = \frac{1}{1-f_1} \cdot \frac{f_1\hat{V}(\hat{\beta}_1)}{\sum_{r=0}^{1} f_r\hat{V}(\hat{\beta}_{r})} + \frac{1}{1-f_0} \cdot \frac{f_0\hat{V}(\hat{\beta}_0)}{\sum_{r=0}^{1} f_r\hat{V}(\hat{\beta}_{r})} = \frac{1}{1-f_1}b_1 + \frac{1}{1-f_0}(1-b_1). \]

Just like in Section 5.1.2 we see that \( k \) is a weighted average of the inverse of the response rates. If all \( f_0 = f \), the overall nonresponse rate, we get that \( k = 1/(1-f) \). Otherwise, a stratum response rate \( 1-f_r \) has large weight if either the nonresponse rate is large and/or the estimated variance of \( \hat{\beta}_r \) is large.

Alternatively, from (17):
\[ \frac{1}{k} = \frac{\sum_{r=0}^{1} (1-f_r)f_r\hat{V}(\hat{\beta}_{r})}{\sum_{r=0}^{1} f_r\hat{V}(\hat{\beta}_{r})} = \sum_{r=0}^{1} a_r(1-f_r) \]
where the weights are \( a_r = f_r\hat{V}(\hat{\beta}_{r}) / \{ f_1\hat{V}(\hat{\beta}_1) + f_0\hat{V}(\hat{\beta}_0) \} \). So we can alternatively express \( 1/k \) as a weighted average of the response rates.

We note that if the aim is to estimate \( \pi_1 \) and \( \pi_0 \) we obtain, of course, \( k = 1/(1-f_1) \) for \( \pi_1 \) and \( k = 1/(1-f_0) \) for \( \pi_0 \).

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 \( x = (x_1, x_2, \ldots, x_{H-1}) \):
\[ \log \frac{P(Y = 1 | x)}{P(Y = 0 | x)} = \alpha + \beta_1x_1 + \beta_2x_2 + \ldots + x_{H-1}\beta_{H-1}. \]
We see that
\[ \alpha = \log \frac{\pi_0}{1 - \pi_0} \]
and
\[ \beta_h = \log \frac{\pi_h}{\pi_0} \left( \frac{(1 - \pi_h)}{(1 - \pi_0)} \right) = \log \text{(odds ratio) for class } h \text{ versus class 0.} \]

Estimating \( \beta_h \) 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 \( s, y = (y_1, ..., y_n) \), 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
\[ P(R_i = 1 \mid y_i = 1) = q_{iy} \text{ and } P(R_i = 1 \mid y_i = 0) = q_{iy}. \]

Otherwise, the model is the same as in Section 5.2 with class probabilities: \( \pi_i = P(Y_i = 1 \mid x_i = 1) \) and \( \pi_0 = P(Y_i = 1 \mid x_i = 0) \), and the logit version
\[ \log \left\{ \frac{P(Y = 1 \mid x)}{P(Y = 0 \mid x)} \right\} = \alpha + \beta x \]
with
\[ \beta = \log \frac{\pi_i}{\pi_0} \left( \frac{(1 - \pi_i)}{(1 - \pi_0)} \right). \]
The aim is still to estimate \( \beta \).

Let now \( s^1 = \{ i \in s : y_i = 1 \} \) and \( s^0 = \{ i \in s : y_i = 0 \} \) with sizes \( n_i^1 \) and \( n_i^0 \). The response samples in the strata are \( s_{ir}^1 = \{ i \in s^1 : R_i = 1 \} \) and \( s_{ir}^0 = \{ i \in s^0 : R_i = 1 \} \) with total response sample being \( s_r = \{ i \in s : R_i = 1 \} = s_{ir}^1 \cup s_{ir}^0 \). The data can now be represented as before, except that nonresponse totals is for each \( y \)-stratum.

<table>
<thead>
<tr>
<th>( x \mid y )</th>
<th>( y = 0 )</th>
<th>( y = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = 0 )</td>
<td>( n_{00r} )</td>
<td>( n_{01r} )</td>
</tr>
<tr>
<td>( x = 1 )</td>
<td>( n_{10r} )</td>
<td>( n_{11r} )</td>
</tr>
<tr>
<td>Totals</td>
<td>( n_{0r}^* )</td>
<td>( n_{1r}^* )</td>
</tr>
<tr>
<td>Nonresponse</td>
<td>( n_{0r}^* - n_{0r} )</td>
<td>( n_{1r}^* - n_{1r} )</td>
</tr>
</tbody>
</table>
The MLE $\hat{\beta}_r, \hat{\pi}_0, \hat{\beta}_l$, 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 $s^l - s_r^l$, the imputed value $x^*$ is drawn at random from $x_{i,obs} = (x_j : i \in s_r^l)$. 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 $n_y^*$ denotes the number of imputed values with $x = i$ and $y = j$:

<table>
<thead>
<tr>
<th>$x \setminus y$</th>
<th>$y = 0$</th>
<th>$y = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = 0$</td>
<td>$n_{00}^*$</td>
<td>$n_{01}^*$</td>
</tr>
<tr>
<td>$x = 1$</td>
<td>$n_{10}^*$</td>
<td>$n_{11}^*$</td>
</tr>
<tr>
<td>Totals</td>
<td>$n_0^* - n_{0r}$</td>
<td>$n_1^* - n_{1r}$</td>
</tr>
</tbody>
</table>

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

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

and

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

We see that

$$
\hat{\beta}_r^l = \log \frac{n_{11r} + n_{11}^*}{n_1 - n_{11r} - n_{11}^*} \quad \text{and} \quad \hat{\beta}_r^0 = \log \frac{n_{01r} + n_{01}^*}{n_0 - n_{01r} - n_{01}^*}.
$$

We now have that conditional on $(y, x_{obs})$, $n_{11}^*$ is binomial $(n_i^* - n_{1r}^*, p^l)$ where $p^l = n_{11r} / n_{1r}$, and $n_{01}^*$ is binomial $(n_i^* - n_{0r}^*, p^0)$ where $p^0 = n_{01r} / n_{0r}$. Then from (16), we find that approximately:

$$
E(\hat{\beta}_r^l \mid y, x_{1,obs}) \approx \log \frac{n_{11r} + (n_i^* - n_{1r}^*)}{n_i - n_{1r} - (n_i^* - n_{1r}^*)} p^l = \log \frac{n_i^* p^l}{n_i^* (1 - p^l)} = \log \frac{p^l}{1 - p^l}
$$

and
Let $f^1$ be the nonresponse rate in stratum $s^1$: $f^1 = (n^*_i - n^*_{ir})/n^*_i$. We note that $\hat{q}_{ir} = n^*_{ir} / n^*_i = 1 - f^1$.

We see that

$$\text{Var}(\hat{\beta}^1_{ir} | y, x_{ir, \text{obs}}) \approx f^1 \left( \frac{1}{n^*_i p^1(1 - p^1)} \right)^2 (n^*_{ir} - n^*_{ir}) p^1(1 - p^1).$$

Similarly,

$$\text{Var}(\hat{\beta}^0_{ir} | y, x_{0, \text{obs}}) \approx f^0 \left( \frac{1}{n^*_0 p^0(1 - p^0)} \right)^2 (n^*_{0r} - n^*_{0r}) p^0(1 - p^0)$$

where $f^0$ is the nonresponse rate in $s^0$: $f^0 = (n^*_0 - n^*_{0r})/n^*_0$. We have that

$$\frac{1}{n^*_{ir} p^1(1 - p^1)} = \frac{1}{n^*_{ir} n^*_i p^1} = \frac{1}{n^*_{ir}} + \frac{1}{n^*_i} \text{ and } \frac{1}{n^*_{0r} p^0(1 - p^0)} = \frac{1}{n^*_{0r}} + \frac{1}{n^*_0}.$$}

So the denominator in (5) becomes

$$E(f^1(1 - f^1)(\frac{1}{n^*_i} + \frac{1}{n^*_r}) + f^0(1 - f^0)(\frac{1}{n^*_{0r}} + \frac{1}{n^*_0})).$$

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

$$E(\hat{\beta}^1_{ir} | y, x_{\text{obs}}) \approx \log \left( \frac{p^1}{1 - p^1} \right) - \log \left( \frac{p^0}{1 - p^0} \right)$$

$$= \log \frac{n^*_i}{n^*_{ir}} - \log \frac{n^*_{0r}}{n^*_0} = \log \frac{n^*_i n^*_0}{n^*_{ir} n^*_{0r}} = \hat{\beta}^1_{ir}.$$}

Hence, the numerator in (5) equals, as before, $\text{Var}(\hat{\beta}^1_{ir}) - \text{Var}(\hat{\beta}^1)$, and exactly as before we have approximately

$$\text{Var}(\hat{\beta}^1_{ir}) - \text{Var}(\hat{\beta}^1) = \frac{1}{n^*_i \pi_i (1 - \pi_i)} \left( \frac{1 - p^1_{ir}}{p^1_{ir}} + \frac{1}{n^*_{0r} \pi_0 (1 - \pi_0)} \frac{1 - p^0_{0r}}{p^0_{0r}} \right).$$

where as before

$$p^1_{ir} = P(R_i = 1 | x_i = 1) \text{ and } p^0_{0r} = P(R_i = 1 | x_i = 0).$$

We need alternative estimates of $p^1_{ir}$ and $p^0_{0r}$:
Since
\[ p_{1r} = P(R_i = 1 | x_i = 1) = P(R_i = 1, Y_i = 1 | x_i = 1) + P(R_i = 1, Y_i = 0 | x_i = 1) \]
\[ = P(Y_i = 1 | x_i = 1)P(R_i = 1 | Y_i = 1) + P(Y_i = 0 | x_i = 1)P(R_i = 1 | Y_i = 0) \]
\[ = \pi_1 q_{1r} + (1 - \pi_1) q_{0r}, \]
we have \( \hat{p}_{1r} = \hat{\pi}_1 (1 - f^1) + (1 - \hat{\pi}_1)(1 - f^0) \) and \( 1 - \hat{p}_{1r} = \hat{\pi}_1 f^1 + (1 - \hat{\pi}_1)f^0. \)

Similarly, \( \hat{p}_{0r} = \hat{\pi}_0 (1 - f^1) + (1 - \hat{\pi}_0)(1 - f^0) \) 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

\[ k = \left( \frac{1}{n_{1r}} + \frac{1}{n_{0r}} \right) (\hat{\pi}_{1r} f^1 + (1 - \hat{\pi}_{1r}) f^0) + \left( \frac{1}{n_{0r}} + \frac{1}{n_{0r}} \right) (\hat{\pi}_{0r} f^1 + (1 - \hat{\pi}_{0r}) f^0) \]
\[ \frac{f^1 (1 - f^1) (\frac{1}{n_{1r}} + \frac{1}{n_{0r}}) + f^0 (1 - f^0) (\frac{1}{n_{0r}} + \frac{1}{n_{0r}})}{f^1 (1 - f^1) (\frac{1}{n_{1r}} + \frac{1}{n_{0r}}) + f^0 (1 - f^0) (\frac{1}{n_{0r}} + \frac{1}{n_{0r}})} \]
\[ = \frac{f^1 (\frac{1}{n_{1r}} + \frac{1}{n_{0r}}) + f^0 (\frac{1}{n_{1r}} + \frac{1}{n_{0r}})}{f^1 (1 - f^1) (\frac{1}{n_{1r}} + \frac{1}{n_{0r}}) + f^0 (1 - f^0) (\frac{1}{n_{0r}} + \frac{1}{n_{0r}})}. \]

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 \( w_1 = \frac{1}{n_{1r}} + \frac{1}{n_{0r}} \) and \( w_0 = \frac{1}{n_{0r}} + \frac{1}{n_{0r}} \). Then

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

where

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

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,

\[ k = \frac{1}{1 - f}, \]  

with \( f = (n - n_r)/n \), the nonresponse rate. \( \quad (20) \)

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

\[
E(k) = \frac{\text{Var}(\hat{\theta}^* | s, s_r) - \text{Var}(\hat{\theta})}{E\text{Var}(\hat{\theta}^* | s, s_r)}. \quad (21)
\]

One obvious requirement is that, at least approximately

\[
E(\hat{\theta}^* | s) = \hat{\theta}, \quad (22)
\]

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

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

\[ \hat{\theta} = \sum_{i \in s} a_i(s) y_i \]  

(23)

Some results:

**Lemma 1** 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., \( \hat{\theta} = a(s) \sum_{i \in s} y_i = na(s) \bar{y}_s \).

**Theorem** Assume \( \hat{\theta} \) is given by (23) and satisfies (22). Then \( E(k) = 1/p_s \) 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, \( \hat{\beta} = \sum_{i \in s} y_i / \sum_{i \in s} x_i \). 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:

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

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

Here one can show that \( E(\hat{\beta}^* \mid s) \approx p_r \hat{\beta} (\text{exact } \frac{n_r - 1}{n - 1} \hat{\beta}) \). 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 \( P(R_i = 1, R_j = 1 \mid s, n_r) = \frac{n_r}{n} \cdot \frac{n_r - 1}{n - 1} \) (follows from (b))
- (d) \( E(Y^*_r \mid s, s_r) = \bar{y}_r \) (\( \Rightarrow E(Y^*_r \mid s, n_r) = \bar{y}_r \Rightarrow E(Y^*_r \mid s) = \bar{y}_s \))
- (e) \( \text{Var}(Y^*_r \mid s, s_r) = \frac{n_r - 1}{n_r} \sigma^2_r \)

\[(\Rightarrow \text{Var}(Y^*_r \mid s, n_r) = \frac{n_r - 1}{n_r} \sigma^2_r \text{, where } \sigma^2_r = \frac{1}{n_r - 1} \sum_{i \in s} (y_i - \bar{y}_s)^2 \text{ and Var}(Y^*_r \mid s) \approx \hat{\sigma}^2_r)\]

**Proof of Lemma 1**

We get

\[ E(\hat{\beta}^* \mid s) = E(\sum_{i \in s} a_i(s) y_i + \sum_{i \in s - s_r} a_i Y^*_i \mid s) \]

\[ = E_{s_r} E(\sum_{i \in s} a_i(s) y_i + \sum_{i \in s - s_r} a_i(s) Y^*_i \mid s, s_r) \]

\[ = (d) E(\sum_{i \in s} a_i(s) y_i \mid s) + E(\sum_{i \in s - s_r} a_i(s) \bar{y}_r \mid s) \]

First term :

\[ E(\sum_{i \in s} a_i(s) y_i \mid s) = E E(\sum_{i \in s} a_i(s) y_i \mid s, n_r) \]

\[ = E E(\sum_{i \in s} a_i(s) y_i R_i \mid s, n_r) = E(\sum_{i \in s} a_i(s) y_i P(R_i = 1 \mid s, n_r)) \]

28
Second term:

\[
E\left[\sum_{i \in x - s} a_i(s) \tilde{y}_r \mid s\right] = E\left[\sum_{i \in x - s} a_i(s) \tilde{y}_r \mid s, n_r\right] \\
= E\left[\frac{1}{n_r} \sum_{i \in x - s} \sum_{j \in s_x} a_i(s) y_j \mid s, n_r\right] \\
= E\left[\frac{1}{n_r} \sum_{i \in x} \sum_{j \in s} a_i(s) y_j (1 - R_i) R_j \mid s, n_r\right] \\
= E\left[\frac{1}{n_r} \sum_{i \in x} \sum_{j \in s} a_i(s) y_j (E(\theta(s) \mid s, n_r) - E(R, R_j \mid s, n_r))\right] \\
= E\left[\frac{1}{n_r} \sum_{i \in x} \sum_{j \in s} a_i(s) y_j \left(\frac{n_r}{n} - \frac{n_r}{n} \frac{n_r - 1}{n - 1}\right)\right] \\
= \frac{1 - p_r}{n - 1} \sum_{i \in x} \sum_{j \in s} a_i(s) y_j = \frac{1 - p_r}{n - 1} (n\bar{a}(s)\bar{y}_r - \hat{\theta}),
\]

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

This implies that

\[
E(\hat{\theta}^\ast \mid s) = p_r \hat{\theta} + \frac{1 - p_r}{n - 1} (n^2 \bar{a}(s)\bar{y}_r - \hat{\theta})
\]

and (22) is equivalent to

\[
 p_r \hat{\theta} + \frac{1 - p_r}{n - 1} (n^2 \bar{a}(s)\bar{y}_r - \hat{\theta}) \Leftrightarrow \hat{\theta}(1 + \frac{1 - p_r}{n - 1} - p_r) = \frac{1 - p_r}{n - 1} n^2 \bar{a}(s)\bar{y}_r
\]

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

and the result follows. \(\square\)

Proof of Theorem

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

\[
E(\hat{\theta}^\ast \mid s, s_r) = a(s)(n_r \bar{y}_r + (n - n_r) \bar{y}_r) = na(s)\bar{y}_r
\]
\[ \text{Var}(\hat{\theta}^* \mid s, s_r) = (a(s))^2 \left( n - n_r \right) \frac{n_r - 1}{n_r} \hat{\sigma}^2_r . \]

Hence,

\[ \text{Var}(\hat{\theta}^* \mid s, s_r) = \text{Var}(na(s)\overline{y}_r) = E\text{Var}(na(s)\overline{y}_r \mid s) + E\text{Var}(na(s)\overline{y}_r \mid s) \]

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

\[ = En^2 [(a(s))^2 \{E_{n,\nu} \text{Var}(\overline{y}_r, s, n_r) + Var\{na(s)E_{n,\nu}E(\overline{y}_r \mid s, n_r)\}\} + Var\{na(s)E_{n,\nu}E(\overline{y}_r \mid s, n_r)\}] \]

\[ = En^2 [(a(s))^2 \{E_{n,\nu} (\frac{1}{n_r} - \frac{1}{n}) + Var\nu_n(\overline{y}_r)\} + Var\{na(s)E_{n,\nu}E(\overline{y}_r \mid s, n_r)\}] \]

\[ = En^2 [(a(s))^2 \{E_{n,\nu} (\frac{1}{n_r} - \frac{1}{n}) + 0 + Var\{na(s)E_{n,\nu}E(\overline{y}_r \mid s, n_r)\}\}] \]

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

Next,

\[ E\text{Var}(\hat{\theta}^* \mid s, s_r) = En^2 [(a(s))^2 \{E_{n,\nu} (\frac{1}{n_r} - \frac{1}{n}) + Var\nu_n\{na(s)E_{n,\nu}E(\overline{y}_r \mid s, n_r)\}\}] \]

\[ = En^2 [(a(s))^2 \{E_{n,\nu} (\frac{1}{n_r} - \frac{1}{n}) + 0 + Var\{na(s)E_{n,\nu}E(\overline{y}_r \mid s, n_r)\}\}] \]

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

We find now, from (21)

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

\[ = \frac{n^2 (E(\frac{1}{n_r} - \frac{1}{n})E[(a(s))^2 \hat{\sigma}^2_r])}{n(1 - pr)E[(a(s))^2 \hat{\sigma}^2_r]} = \frac{n(E(1/n_r) - 1)}{1 - pr} \approx \frac{(1/pr) - 1}{1 - pr} = \frac{1}{pr} . \]

\[ \square \]

Reference

Recent publications in the series Discussion Papers

332 M. Greanker (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
340 H. C. Bjørnland and H. Hungnes (2003): The importance of interest rates for forecasting the exchange rate
343 B. Bye, B. Strom and T. Ávísland (2003): Welfare effects of VAT reforms: A general equilibrium analysis
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
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
352 E. Holmøy (2003): Aggregate Industry Behaviour in a Monopolistic Competition Model with Heterogeneous Firms
369 T. Skjerpen (2004): The dynamic factor model revisited: the identification problem remains