Weighted Hotdeck Imputation

Adrian Mander

MRC Biostatistics Unit, Cambridge, UK.

Corresponding author: adrian.mander@mrc-bsu.cam.ac.uk

David Clayton

MRC Biostatistics Unit, Cambridge, UK.

Summary.
Hotdeck imputation replaces missing lines of data by sampled complete data records. The theory underpinning this method was published over a decade ago (Rubin and Schenker1986), and since its conception has remained predominantly unchanged. Most of the research involving hotdeck imputation has been in the area of constructing strata, which are used to impute within (independently), notably the propensity score (Rosenbaum and Rubin1984). This paper extends these ideas by using the actual propensity score in a weighted hotdeck imputation procedure. The weighted hotdeck is compared to the original hotdeck using simulated data examples and a real data example previously covered in (Pickles *et al.*1995). The real data example demonstrates an application of multiple imputation when the missing data is by design.

Keywords: missing data, Multiple Imputation, multi-phase designs, hotdeck imputation, approximate bayesian bootstrap

1. Introduction

An imputation method is one that replaces missing data by imputed values to create a complete dataset. Standard methods are then used to analyse the complete dataset avoiding the use of complex missing data likelihoods. The imputation method is applied to a dataset several times creating several complete datasets, as part of a multiple imputation scheme. Each dataset is analysed using the same statistical method and the parameters of interest can be combined, according to the formulae specified in (Rubin1987). Hotdeck imputation is simple with very few distributional assumptions, but its application within mainstream statistics is not widespread although the theoretical basis was developed over a decade ago (Rubin and Schenker1986). Hotdeck imputation assumes that the missing data are ignorable, either missing at random (MAR) or missing completely at random (MCAR) (Little and Rubin1987). Several statistical packages treat missing data as MCAR by default and use case-only analysis, this can lead to bias when the missing data are not MCAR(Wang-clow *et al.*1995).

Hotdeck imputation is a semi-parametric method which only uses the observed empirical distribution of the data and makes no other distributional assumptions. Other imputation methods for both univariate and multivariate missing can be found in (Rubin1987; Schafer1997).

Hotdeck imputation can be used to handle monotone patterns of missingness (Lavori *et al.*1995) and when dealing with large datasets a more efficient implementation is discussed in (Reilly1993). The hotdeck estimator has the same asymptotic distribution as the mean score estimator

(Reilly and Pepe1996) and has strong links to weighted likelihood (Reilly and Pepe1997). Hotdeck imputation has been implemented for the statistical package STATA (Mander and Clayton1999).

This paper investigates a version of hotdeck called the Approximate Bayesian Bootstrap (ABB) (Rubin and Schenker1986; Rubin1987). ABB imputes missing data by sampling from the observed data, the detailed method is in section 2. To improve accuracy sampling can be stratified by variables that predict whether the data are missing. Selection of strata to impute within can be made using a model-based measure such as the propensity score (Rosenbaum and Rubin1983). Stratification based on the propensity of missing has been used with hotdeck imputation (Rosenbaum and Rubin1984). The propensity score is constructed using logistic regression, the fitted probabilities are categorised into strata and imputation is performed within strata independently.

Weighted hotdeck imputation uses the actual propensity score, rather than the "discrete" strata, to weight the sampling of observed data. Therefore, the weighted hotdeck uses more information from the missing data mechanism, assuming that this model is accurate, and should be closer to the true complete dataset.

The weighted hotdeck will be illustrated by using data taken from a multi-phase study design (Pickles *et al.*1995). This type of design is not usually analysed using multiple imputation. Simulation models are also used to compare the weighted hotdeck to the original hotdeck and to check the coverage of the confidence intervals.

2. Weighted Hotdeck Imputation

Hotdeck imputation has been used as a method for imputing values in a single variable. Applying the method when missing data are univariate but analysis has been multivariate has led to biased results (Allison2000). The example in (Allison2000) is a regression analysis of Y against X, where X has missing values. Sampling from observed values of X leads to regression dilution bias. This is due to the imputing mechanism destroying the relationship between X and Y. A possible solution is by making a parametric assumption about the distribution $X|Y$. However the hotdeck imputation model can be applied by imputing (X, Y) pairs of data which contain at least one missing datum. Some information is lost when observed values in the lines with missing data are replaced, but the observed relationships among the variables remain unchanged.

Describing hotdeck imputation more formally; let a dataset $\mathbf{Z}$ consist of i variables $(Z_1, Z_2, \ldots, Z_i)$ each with n elements. The complete rows of data can be represented by $\mathbf{Z}_{com}$ and let there be $n_1 (< n)$ observed rows. Similarly, the incomplete data $\mathbf{Z}_{inc}$ are the n_0 (where $n = n_1 + n_0$) rows of data that contain missing values for the variables Z_k where $k = j, \ldots, i$, for some $j > 1$. This dataset is represented in figure 1 where the rectangles represent observed data. To incorporate other patterns of missing into the same framework let $\mathbf{Z}_{inc}$ be the rows of data that contain *at least one* missing value for the variables Z_k where $k = j, \ldots, i$. In the imputation the variables Z_k $(k = j, \ldots, i)$ are considered missing for the first n_0 elements.

An imputation model creates a full dataset by drawing from the posterior distribution of $\mathbf{Z}_{inc}$. With the same model m draws are made as part of a multiple imputation scheme, these m complete datasets are analysed separately to give a series of parameter estimates, $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots, \hat{\theta}^{(m)}$. These parameter vectors are combined to give an overall estimate.

A basic imputation model is Simple Random Imputation (SRI); where the n_0 rows of $\mathbf{Z}_{inc}$ are randomly sampled with replacement, from $\mathbf{Z}_{com}$ (Rubin and Schenker1986).

The resulting confidence intervals of $\hat{\theta}$ are too narrow (Rubin1987) because the parameters of the empirical distribution $\mathbf{Z}_{com}$ are assumed to be the population parameters. Hence there is a need to sample the parameters of the empirical distribution of $\mathbf{Z}_{com}$ before drawing from the posterior of $\mathbf{Z}_{inc}$. This methodology applied to the SRI model led to the Bayesian Bootstrap (BB) and the Approximate Bayesian Bootstrap (ABB) (Rubin and Schenker1986; Rubin1987). The ABB is represented in figure 1. The first step is the sampling with replacement of $\mathbf{Z}_{com}$ to create $\mathbf{Z}^*_{com}$, note $\mathbf{Z}^*_{com}$ has the same dimensions as $\mathbf{Z}_{com}$. Secondly n_0 rows are sampled with replacement from $\mathbf{Z}^*_{com}$ to create the imputed dataset $\mathbf{Z}_{imp}$.

The methods can be adapted if the missing data mechanism is known to vary between various strata. For example, take two equal sized stratum: in one the probability of missing data is 0.5; and in the second stratum the probability of missing data is 0.1. This example could arise from missing by design studies, such as multi-phase studies. By ignoring the strata missing data are more likely to be replaced by data in the second stratum using hotdeck imputation. The solution is to use hotdeck imputation within each stratum separately. A major flaw of this method is that as the number of strata increase, imputation within stratum becomes impractical. In the extreme case imputation will fail when one stratum arises that contains only missing data and no observed data.

Identification of a reasonable number of strata can be made using the propensity score (Rosenbaum1998). Consider a study comparing two treatments where $V = 1$ or 0 is the treatment assignment. The propensity score, e, is the estimated conditional probability that a unit with vector W of observed covariates will be assigned to treatment 1, $e(W) = P(V = 1|W)$, with the property that W and V are conditionally independent given $e(W)$. The propensity score is estimated via logistic regression with W as the covariates (Rosenbaum and Rubin1984), (interaction terms may be needed). The propensity score is categorised into as few as 5 classes (Rosenbaum and Rubin1984). Using the hotdeck within these subclassifications is intuitively appealing, as missing data will be replaced by data that are close, as measured by the propensity score (Lavori *et al.*1995). Here the treatment variable V is replaced by a variable indicating missing data. The different patterns of missing data need to be included within the set of covariates.

2.1. Sampling weights

An indicator variable Δ is constructed where $\Delta = 1$ if a row of $\mathbf{Z}$ contains missing data and 0 otherwise.

Using Bayes theorem,

$$\begin{aligned}
[\mathbf{Z}_{inc}] &= [Z_1, \ldots, Z_{j-1}|\Delta = 1], \\
&= [Z_1, \ldots, Z_i|\Delta = 0]\frac{[\Delta = 1|Z_1, \ldots, Z_{j-1}]}{[\Delta = 0|Z_1, \ldots, Z_i]}\frac{[\Delta = 0]}{[\Delta = 1]}, \\
&\propto [Z_1, \ldots, Z_i|\Delta = 0]\frac{[\Delta = 1|Z_1, \ldots, Z_{j-1}]}{[\Delta = 0|Z_1, \ldots, Z_i]}, \\
&= [Z_1, \ldots, Z_i|\Delta = 0]\frac{[\Delta = 1|Z_1, \ldots, Z_{j-1}]}{[\Delta = 0|Z_1, \ldots, Z_{j-1}]}, \qquad (1)
\end{aligned}$$

where [ ] indicates a probability distribution or density function. The $Z_j, \ldots, Z_i$ variables are removed from the right-hand side condition in the last line because the missing data mechanism is assumed to be ignorable. The term $\frac{[\Delta=1|Z_1,\ldots,Z_{j-1}]}{[\Delta=0|Z_1,\ldots,Z_{j-1}]}$ is the odds of missing

given $Z_1, \ldots, Z_{j-1}$. These can be estimated by logistic regression and provide the sampling weights of the joint distribution of the complete rows of data ($\mathbf{Z}_{com}$). The constant term in this regression is $\frac{[\Delta=0]}{[\Delta=1]}$ but, as the weights are constrained to sum to 1, this term is not required.

In the ABB the parameters are sampled before drawing from the posterior. In the weighted hotdeck the parameters are the odds ratio of missing and the empirical distribution of the complete lines of data. Firstly, one could perturb the vector of regression coefficients (in the logistic regression to predict missing) using the estimated covariance-variance matrix. However, an alternative method suggested by the ABB is illustrated in figure 2. $\mathbf{Z}_{inc}$ is sampled with replacement to form $\mathbf{Z}^*_{inc}$ and $\mathbf{Z}_{com}$ is sampled with replacement to form $\mathbf{Z}^*_{com}$. The conditional sampling means that the probability of missing data is unaffected. $Z^*_1, \ldots, Z^*_{j-1}$ is used in the logistic regression to predict the odds ratio ($\frac{[\Delta=1|Z^*_1,\ldots,Z^*_{j-1}]}{[\Delta=0|Z^*_1,\ldots,Z^*_{j-1}]}$). Using $Z^*_1, \ldots, Z^*_{j-1}$ in predicting the odds ratio rather than $Z_1, \ldots, Z_{j-1}$ is equivalent to resampling the odds ratio. The final step of the weighted hotdeck is to do a weighted sampling with replacement of $\mathbf{Z}^*_{com}$ using the perturbed odds ratio to form $\mathbf{Z}_{imp}$.

With the imputation model a series of complete datasets are constructed and used as part of a multiple imputation framework. Results from the analysis of the completed dataset are combined using the rules described by (Schafer1997; Rubin1987). The number of imputations needed is indicated by the relative efficiency $(1 + \lambda/m)^{-1}$, where λ is the fraction of missing information and m the number of imputations. Unlike Gibbs Sampling very few imputations are actually needed (≈ 20). More imputations, than obtained by the formula, may be needed when the between imputation variance is larger than the within variance and in non-standard datasets.

3. Simulation Study

Data will be generated using three simulation models. The first two models share the following in common.

The dependent variable Y and exposure X are both binary (1/0). The distribution of X is

$$X = \begin{cases} 1 & \text{w.p. } p_{x1} \\ 0 & \text{o.w.} \end{cases} .$$

In our simulations p_{x1} is 0.5. The Y variable is constructed using a logistic regression model. Given the two regression coefficients are α and β ,

$$P(Y = 1/X) = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)},$$

for our simulations $\alpha = \log(0.5)$ and $\beta = \log(2)$. Y is constructed by taking observations u_i from a uniform(0,1) distribution such that

$$Y_i = \begin{cases} 1 & \text{if } u_i < P(Y = 1|X) \\ 0 & \text{o.w.} \end{cases} .$$

3.1. Missing Data Mechanism: Model 1

Model 1 will generate MAR data so that ignoring the missing data will lead to biased results. This is achieved by using differential probabilities of missing in each stratum of X.

Typically the distribution of missing data takes the following form.

$$P(Y \text{ is missing}) = \begin{cases} p_{m1} & \text{if X=1} \\ p_{m0} & \text{if X=0} \end{cases} .$$

If $p_{m1} = p_{m0}$ then the missing data mechanism is MCAR. If these differ then the missing data mechanism depends on X but not Y.

3.2. Missing Data Mechanism: Model 2

Here the missing data mechanism depends on a variable that is not used in the analysis. Taking the artificial example of a variable with distribution

$$G = \begin{cases} 1 & \text{if X=1 } \textbf{and} \text{ Y=1} \\ 0 & o.w. \end{cases} .$$

If the missing data depended on G, but G was unobserved, then the missing data mechanism would be non-ignorable, since it depends on the outcome Y. In reality this is a trick to construct data that displays missing data that on the surface is non-ignorable however a variable observed on everyone allows the assumption of MAR data. For the simulations a surrogate of G, $\tilde{G}$ will be used. The surrogate is constructed stochastically with sensitivity, $P(\tilde{G} = 1|G = 1)$, and specificity, $P(\tilde{G} = 0|G = 0)$ equal to p_{s1} and p_{s2} respectively. In our simulations the surrogate is very good, p_{s1} and p_{s2} are both 0.9. The missing data mechanism is,

$$\text{P(Y is missing)} = \begin{cases} p_{m1} & \text{if } G = 1 \\ p_{m0} & \text{if } G = 0 \end{cases} .$$

3.3. Missing Data Mechanism: Model 3

The simulation model described below has been used in several papers (Robins *et al.*1994; Pepe and Fleming1991; Reilly1993).

To obtain a generated dataset; the following are simulated for each of n subjects, a normally distributed exposure $X \sim N(0, 1)$, a dichotomous surrogate $G = I[X + \nu > 0]$, where $\nu \sim N(0, \sigma^2)$, ν is independent of X and $I[A] = 1$ if A is true and 0 otherwise. The binary outcome is generated from the logistic probability function $P(Y = 1|X = x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$. The simulation study will investigate estimates of $Var(\hat{\beta})$ when X has missing data that is MCAR. The probability of a missing value in X is 0.2, to allow comparison to the results of the following papers (Robins *et al.*1994; Pepe and Fleming1991; Reilly1993). The question of whether the surrogate improves variance estimates can be investigated by varying the σ parameter. As the variance of ν increases the informativeness of the surrogate of G decreases.

3.4. Investigation of bias in Models 1 and 2

Simulation model 1 and 2 are used to generate a dataset of size 20000, with $p_{m0} = 0.4$, $p_{m1} = 0.1$ and the specificity and sensitivity equal to 0.9 for model 2. The results using both hotdeck methods with 30 imputations are displayed in Table 1. From the table it is clear that the estimates' confidence intervals all contain the true values ($\alpha = -\log(2)$ and $\beta = \log(2)$). There is a suggestion that the weighted hotdeck results have slightly larger standard errors for the constant parameter.

3.5. Assessing Coverage

It is possible to investigate any bias in the imputation variance estimate using simulation models. In the simulated datasets the true population parameter of interest, θ say, is known.

To help with the terminology in this section, the formula for combining m estimates of θ $(\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots, \hat{\theta}^{(m)})$ from the m imputed datasets is outlined below (Rubin1987). Let

$$\overline{\theta}_m = \sum_{l=1}^{m} \hat{\theta}^{(m)}/m,$$

be the average of the m complete-data estimates,

$$\overline{U}_m = \sum_{l=1}^{m} Var(\hat{\theta}^{(m)})/m,$$

be the average of the m complete-data variances, and

$$B_m = \sum_{l=1}^{m} (\hat{\theta}^{(m)} - \overline{\theta}_m)^2/(m-1),$$

be the variance between the m complete-data estimates. The total variance is defined as

$$\overline{U}_m + (1 + m^{-1})B_m. \tag{2}$$

In order to obtain the bias in the coverage, take $\hat{W}$ as the complete data variance estimate of θ and $\hat{B}$ to be the estimate of the variance between imputations. Assuming that an infinite number of imputations are used to estimate the variance of θ, then this estimate is $\hat{V} = \hat{W} + \hat{B}$ (from equation 2). Let V be the true estimate of this variance. Then take S as the variance of a single imputation estimate about θ (if θ is unknown then $\hat{S}$ is also unknown), then $\hat{S} = \hat{B} + V$. It follows that the bias in $\hat{V}$ is $\hat{V} - V = W - S + 2B$.

Estimates for these variances are found by the following steps; the imputation model is used with 2 imputations, this is the minimum to obtain an estimate of the between variance estimate. From the two complete datasets two estimates of θ are obtained, $\hat{\theta}_1$ and $\hat{\theta}_2$ with the corresponding variance estimates $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$. Pairs of imputations are repeated N times to obtain estimates of the various variance components. From the imputations the estimates for $\hat{S}, \hat{B}, \hat{W}$ are given below, using i as the subscript for the number of iterations $(i = 1, 2, \ldots, N)$:

$$\begin{aligned}
\hat{S} &= 1/2N \sum_i (\hat{\theta}_{1i} - \theta)^2 + (\hat{\theta}_{2i} - \theta)^2, \\
\hat{B} &= 1/2N \sum_i (\hat{\theta}_{1i} - \hat{\theta}_{2i})^2, \\
\hat{W} &= 1/2N \sum_i (\hat{\sigma}_{1i}^2 + \hat{\sigma}_{2i}^2).
\end{aligned}$$

From these the bias and the relative bias $\frac{\hat{V}-V}{V}$ can be obtained.

Data are simulated from model 1 and 2 of sizes 2000, 4000 and 8000. The probabilities of missing data p_{m1} and p_{m0} have been arbitrarily been selected with the condition $p_{m1} > p_{m0}$, this is missing by design, where exposed subjects are more likely to enter the analysis

for power reasons. This design is usually advantageous when exposure is rare. Here the parameter of interest is the log odds ratio of being a case comparing exposed and unexposed subjects. The number of times the process has been repeated is 2000 ($= N$). The relative bias, $(\hat{V} - V)/V$ and the average of the $\hat{V}$'s are given in table 2.

The average of the $\hat{V}$'s for hotdeck imputation are nearly identical to the weighted hotdeck estimates. For model 1 the range of estimates of the relative bias is (-6.5% to 3.2%) for hotdeck and (-4.2% to 6.3%) for the weighted hotdeck. The relative bias is indicating that both methods appear to be unbiased although there are some fluctuations due to simulation error. For model 2 the range of the relative bias is (-10.2% to 7.4%) for hotdeck and (-10.1% to 16.1%) for the weighted hotdeck. Again the relative bias is indicating that both methods give correct coverage. The simulation error appears to be larger in model 2 than in model 1 but this is mainly due to the uncertainty of the surrogate. It is clear that the models can handle quite extreme probabilities of missing and give reasonable estimates of the variance. The relative bias does not seem to depend on the probabilities of missingness or the sample size of the dataset.

3.6. Do surrogates give extra information when data are MCAR

When the missing data mechanism is MCAR a complete-case analysis will give unbiased (inefficient) estimates of the parameters of interest. Hotdeck imputation is an ideal method for this situation, when the full dataset is not large. Simulation model 3 is used to investigate small datasets when the missing data are MCAR and a surrogate of the missing variable is available.

Hotdeck imputation was used to analyse this simulation model in (Reilly1993), here the variance of β is estimated using an adaptive hotdeck. In addition to the adaptive hotdeck the simulation study has a balanced design of missing data, in other words there are equal numbers of complete data in each of the stratum defined by G and Y, this improves the variance estimates. Using the same study design and the parameters used in (Reilly1993) weighted hotdeck and hotdeck are used to obtain estimates of this model and are displayed in tables 3 and 4. The variance estimates are based on a sample size of 4000, as the between variance estimate requires 2 simulations and is based on a sample size of 2000.

The weighted hotdeck relative efficiencies are close to 1 when $\sigma^2 = 2$ but when σ^2 is closer to 0 the relative efficiency is reduced to about 0.6. This reduction of efficiency is not observed in hotdeck imputation, see table 4. When the missing data are MCAR the logistic regression to predict the probability of missing should have coefficients that are, in expectation, 0. In reality there are some gains in the efficiency due to capturing some of the sampling variation in the estimates. The hotdeck imputation uses the information of the surrogate directly by sampling within strata of the surrogate and is unaffected by the missing data mechanism. It is interesting to note that the relative efficiency of the weighted hotdeck is not lost when the surrogate deteriorates.

The second set of results shown in tables 5 and 6 use the same parameters and design as the paper (Pepe and Fleming1991). The missing data are not balanced and hence the variance estimates will be larger. The estimator in (Pepe and Fleming1991) is extremely efficient and its asymptotic relative efficiency comparing it to the most efficient estimator for this study design is given in (Robins *et al.*1994). As the results in (Robins *et al.*1994) are displayed in terms of asymptotic relative efficiencies and the variances are not reported it is difficult to compare the hotdeck and weighted hotdeck estimators with the most efficient estimator. However, the Pepe-Flemming estimator (Pepe and Fleming1991) achieves an

ARE of 1 when $\beta = 0$ in the simulation model and this degrades as β increases.

As can be seen from table 5 the weighted hotdeck achieves around 0.3 relative efficiency when the sample size is 200 and 0.1 when the sample size is 100. The results in table 6 for hotdeck are slightly better with relative efficiencies of around 0.45 and 0.2, respectively. Table 5 shows that when the surrogate is good efficiency is lost using the weighted hotdeck. The hotdeck methods would be more viable when the sample size of the complete dataset was larger. Also estimation would be improved by reducing the probability of missingness.

4. Two-Phase Study Example

When missing data are by design there are definable strata to include in the imputation models. This does assume that the missing data only resulted from the design, it is possible to have more than one missing data mechanisms present. Missing by design is one area that imputation methods can greatly simplify analysis, these designs are becoming increasingly used in epidemiological circles, for example, the multi-phase design (White1982). This design involves collecting data on the whole study population, this could be demographics or surrogate measures of covariates (or outcome variables if the covariates have missing data). The next stage involves collection of detailed data on a subsample of the first phase population, for example, using more expensive laboratory tests. An example of this type of design is covered in (Pickles *et al.*1995). Here patients' mental health was initially assessed by a general practitioner (GP) and the Spanish version of the 28-item general health questionnaire (GHQ). A patient was classified as a case (SCREEN=1) if one of the tests were positive (GP=1 or GHQ=1), and a control (SCREEN=0) if both tests were negative (GP=0 and GHQ=0). This first phase was a cross-sectional survey. From the screening phase half the patients with a positive screen were sampled into the second phase and 10% from the negative screens. The majority of screens were negative hence the smaller sampling fraction (maximising information). The second phase sample were assessed by a senior psychiatrist using the Spanish version of the Schedules for Clinical Assessment in Neuropsychiatry (SCAN), this is considered the gold standard. The paper analyses this data using conditional probabilities, bayesian estimation using Gibbs sampling (Spiegelhalter *et al.*1994), estimation using expansion weights, weighted logistic regression, an independence working model, generalised linear mixed models and EM approaches for log-linear models. The results from these methods can be compared to the multiple imputation results using the same data, the analysis is displayed in table 7. One distinct advantage of multiple imputation over the other methods is that since the imputed datasets contain no missing data complete data methods are used to obtain prevalence estimates. The robust standard errors and Taylor series expansions to obtain estimates of the standard errors are not needed.

For each of the imputation models the prevalence is estimated by sex (π_m for males and π_f for females). These probabilities will have binomial variance using N_m,N_f and N as the sample sizes of males, females and total, respectively. The data are MAR as the sampling scheme varied between the two levels of SCREEN. The missing data mechanism is predicted by using SCREEN as a predictor in the weighted hotdeck model and as strata in hotdeck. There is some sampling variation between SEX, GP and GHQ and the model could contain all these variables in the model to predict missingness. A table in (Pickles *et al.*1995) investigates the various strata using three sets of expansion weights: weights varying within SCREEN (2 strata); weights varying by SEX and SCREEN (4 strata) and; weights varying

by GP,GHQ and sex (8 strata, SCREEN is defined by GP and GHQ). The same strata are used for the hotdeck and in the weighted hotdeck.

The number of imputations needed is based on the calculations disclosed in (Rubin1987), for 97% efficiency around 20 imputations are needed. The dataset contains 75% missing data and the between imputation variance was greater than the within imputation variance. The number of imputations are therefore taken as a nominal 100 imputations.

From table 7 the strata do not change the estimates of the prevalences' standard errors or point estimates. The point estimates are essentially the same as the analysis in (Pickles *et al.*1995). However, there is a slight reduction of the standard error in the prevalence of female psychopathy when compared to the expansion weight analysis and the conditional probability method. The results have slightly higher standard errors than the Gibbs sampling method, this is the gain due to complexity of the modelling. When using an expansion weights analysis for the various strata the method can be improved by allowing the smoothing of the weights using a technique called 'raking' (Deming and Stephan1940). Essentially this is achieved by using a model based estimate of the weights, this is the method used by weighted hotdeck imputation.

5. Discussion

This paper highlights the need to collect covariates that may predict missingness. Identification of these variables are straightforward when the missing data are by design. However, there are gains in efficiency when including variables that are not really related to missingness (Robins *et al.*1994). Additionally there may be missing data by design in conjunction with other missing data mechanisms which suggests that understanding the missing data mechanism is a necessity before using the imputation methods.

Given an understanding of the missing data which hotdeck method is preferable? From the simulation models the hotdeck imputation gave lower standard errors. Both methods are fairly easy to implement in any statistical package. Weighted hotdeck imputation is more flexible than hotdeck imputation and in cases where numerous strata are needed to explain the missing data mechanism hotdeck imputation may be impossible to implement.

In light of the increase use of the multi phase study design in epidemiological studies the analyst should remember that imputation has a role in estimating the parameters of interest. For many multi phase designs inverse probability weights are generally the method of choice an example is illustrated in the genetic epidemiology literature (Whittemore and Halpern1997; Whittemore1997). The drawback of the weighted analysis is that in non-standard analyses analysis requires specialist software or skills in programming. Multiple imputation uses complete data methods and can be used in standard statistical packages and the standard errors are estimated with relative ease.

The results have been illustrated using simulation models that have univariate missing patterns. The methods can be used when missing data are present in multiple variables. When using weighted hotdeck the various multivariate missing patterns should be included as indicator variables in the logistic regression to predict missingness. Application of hotdeck and weighted hotdeck imputation should be used with caution to tackle multivariate missing data situations, as the efficiency of the methods can drop considerably. These methods replace missing lines of data to preserve the inter-relationships of variables in the dataset, however when there is multivariate missing this will lead to the deletion of some information when the lines of data are not completely missing. As the imputation models do not make

multivariate distributional assumptions these methods will be more robust.

References

[Allison2000] Allison, P. (2000). Multiple imputation for missing data: a cautionary tale. *Sociological Methods and Research*, **28**, 301–9.

[Deming and Stephan1940] Deming, W. and Stephan, F. (1940). On a least squares adjustment of a sample frequency table when the expected margin totals are known. *Annals of Mathematical Statistics*, **11**, 427–44.

[Lavori *et al.*1995] Lavori, P., Dawson, R., and Shera, D. (1995). A multiple imputation strategy for clinical trials with truncation of patient data. *Stat. Med.*, **14**, 1913–25.

[Little and Rubin1987] Little, R. and Rubin, D. (1987). *Statistical analysis with missing data.* J. Wiley & Sons.

[Mander and Clayton1999] Mander, A. and Clayton, D. (1999). Hotdeck imputation. *Stata Technical Bulletin*, **51**, 32–4.

[Pepe and Fleming1991] Pepe, M. and Fleming, T. (1991). A nonparametric method for dealing with mismeasured covariate data. *J.A.S.A.*, **86**, 108–13.

[Pickles *et al.*1995] Pickles, A., Dunn, G., and Vazquez-Barquero, J. (1995). Screening for stratification in two-phase ('two-stage') epidemiological surveys. *Statistical Methods in Medical Research*, **4**, 73–89.

[Reilly1993] Reilly, M. (1993). Data analysis using hot deck multiple imputation. *The Statistician*, **42**, 307–13.

[Reilly and Pepe1996] Reilly, M. and Pepe, M. (1996). A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, **82**, 299–314.

[Reilly and Pepe1997] Reilly, M. and Pepe, M. (1997). The relationship between hot-deck multiple imputation and weighted likelihood. *Stat. Med.*, **16**, 5–19.

[Robins *et al.*1994] Robins, J., Rotnitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *J.A.S.A.*, **89**, (427), 846–66.

[Rosenbaum and Rubin1983] Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

[Rosenbaum and Rubin1984] Rosenbaum, P. and Rubin, D. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J.A.S.A.*, **79**, (387), 516–24.

[Rosenbaum1998] Rosenbaum,P.R. (1998). Propensity Score. *in Encyclopedia of Biostatistics*, J. Wiley & Sons (Chichester).

[Rubin1987] Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* J. Wiley & Sons (New York).

[Rubin and Schenker1986] Rubin, D. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J.A.S.A.*, **81**, 366–74.

[Schafer1997] Schafer, J. (1997). *Analysis of incomplete multivariate data.* Chapman and Hall(London).

[Spiegelhalter *et al.*1994] Spiegelhalter, D., Thomas, D., Best, N., and Gilks, W. (1994). *Bugs manual 0.30.* MRC Biostatistics Unit, Cambridge.

[Wang-clow *et al.*1995] Wang-clow, F., Lange, M., Laird, N., and Ware, J. (1995). A simulation study of estimators for rates of change in longitudinal studies with attrition. *Stat. Med.*, **14**, 283–97.

[White1982] White, J. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *Am. Journal of Epidemiology*, **115**, 119–28.

[Whittemore1997] Whittemore, A. (1997). Multistage sampling designs and estimating equations. *J.R.S.S. B*, **59**, 589–602.

[Whittemore and Halpern1997] Whittemore, A. and Halpern, J. (1997). Multi-stage sampling in genetic epidemiology. *Stat. Med.*, **16**, 153–67.

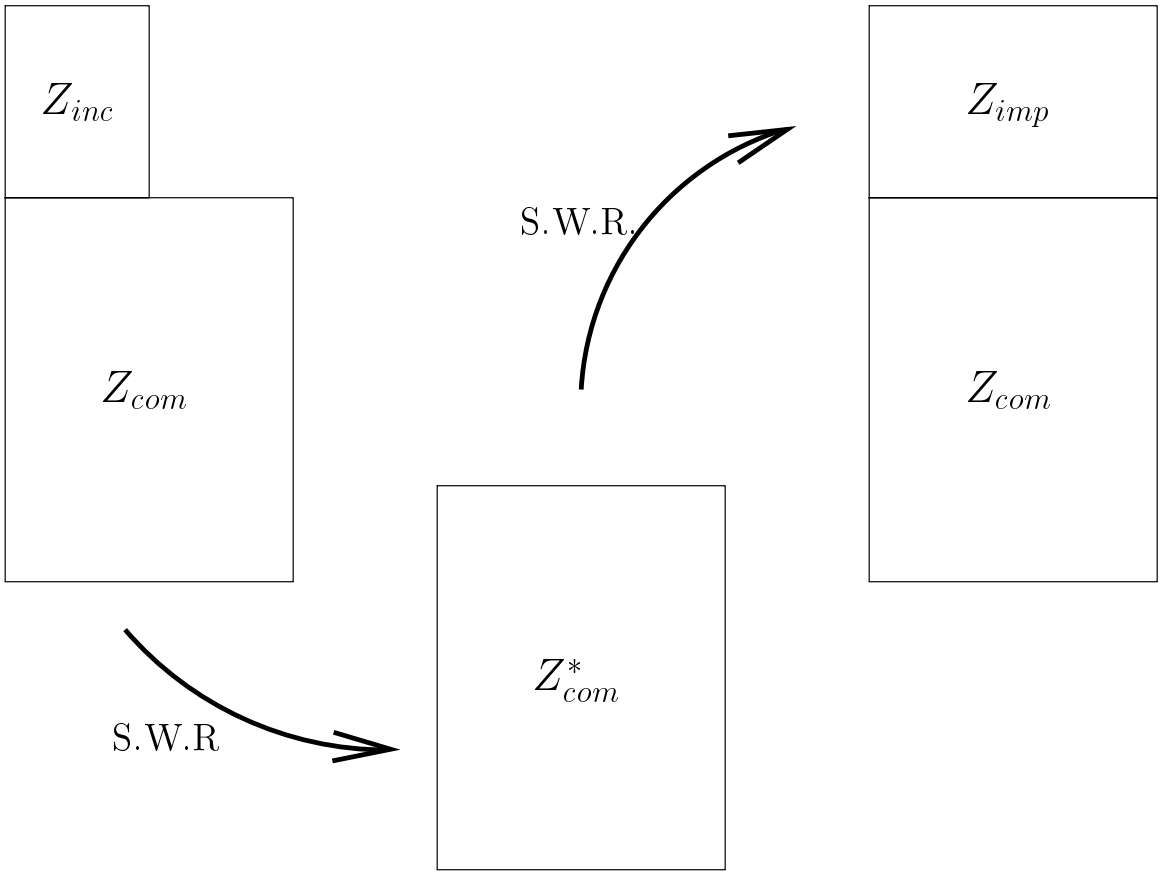

Fig. 1. The Approximate Bayesian Bootstrap

a

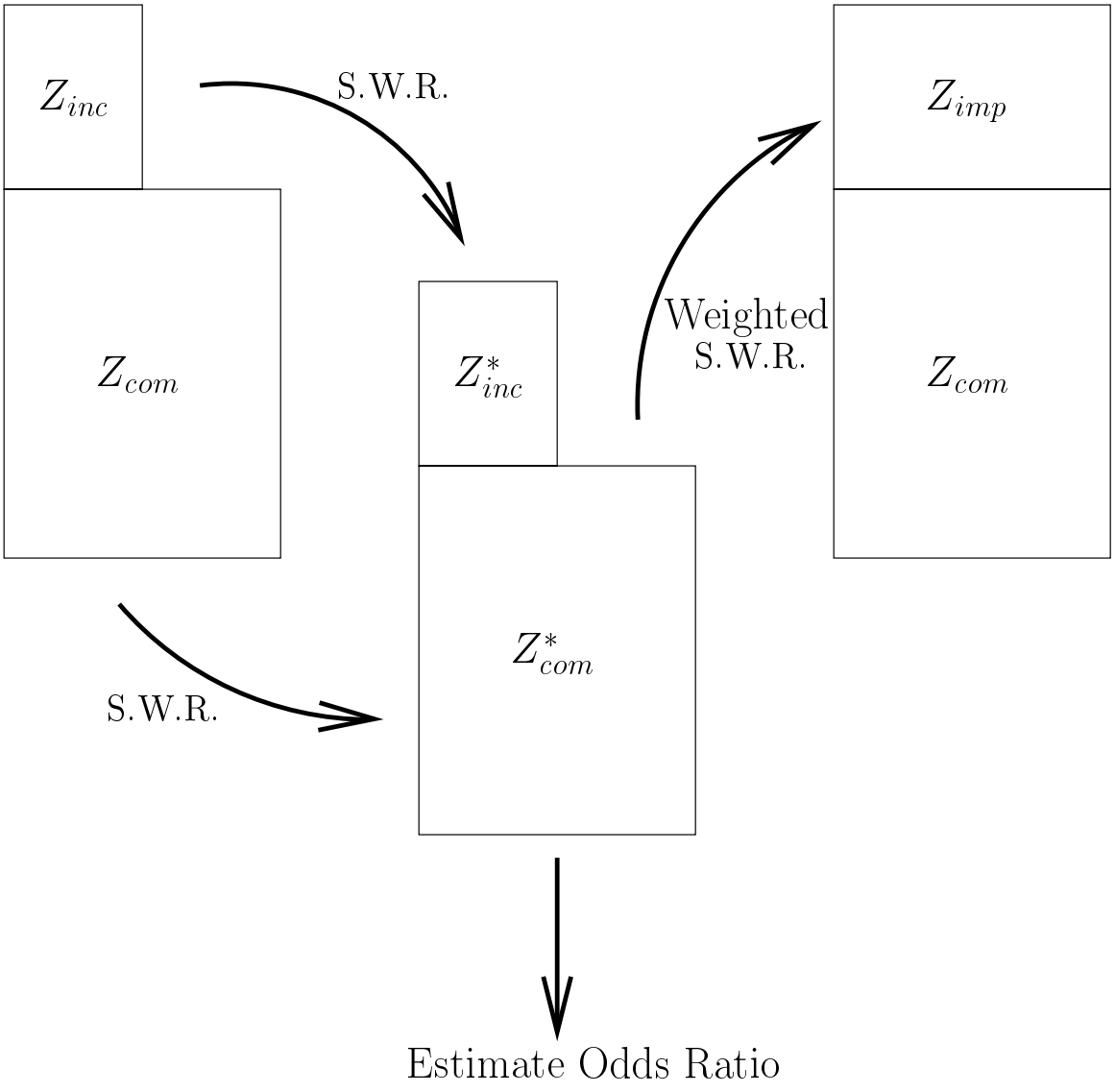

Fig. 2. Adapting the Approximate Bayesian Bootstrap to include Weighted Sampling

Table 1. Investigating potential bias in the logistic regression parameters of the simulated case-control study.

	Method	$\hat{\alpha}$ (SE)	$\hat{\beta}$ (SE)
model 1	Weighted Hotdeck	.724 (.036)	-.699 (.022)
	Hotdeck	.723 (.032)	-.698 (.022)
model 2	Weighted Hotdeck	.659 (.032)	-.683 (.023)
	Hotdeck	.656 (.031)	-.681 (.024)

Table 2. Relative bias in the estimated variance of the log odds ratio

			Model 1:Relative bias ($\hat{V}$)		Model 2: Relative bias ($\hat{V}$)	
Dataset size	$m0$	$m1$	hotdeck	weighted hotdeck	hotdeck	weighted hotdeck
2000	.1	.4	.032 (.006)	.05 (.006)	.02 (.006)	-.015 (.006)
2000	.2	.4	.005 (.006)	.063 (.006)	.06 (.006)	.051 (.007)
2000	.1	.6	-.032 (.008)	-.003 (.008)	-.047 (.009)	.045 (.009)
2000	.4	.6	-.006 (.009)	-.04 (.009)	.062 (.009)	.161 (.01)
2000	.6	.6	-.007 (.01)	.009 (.011)	-.026 (.009)	.08 (.011)
2000	.1	.9	-.011 (.026)	-.042 (.025)	-.079 (.03)	-.093 (.031)
2000	.2	.9	0 (.027)	.018 (.027)	-.102 (.03)	-.101 (.031)
2000	.4	.9	-.065 (.026)	-.015 (.027)	.063 (.032)	.005 (.033)
2000	.6	.9	.015 (.029)	-.018 (.028)	.036 (.031)	-.036 (.031)
4000	.6	.9	.05 (.014)	-.018 (.014)	-.078 (.015)	-.034 (.016)
8000	.6	.9	.011 (.007)	.004 (.007)	.074 (.008)	-.052 (.008)

Table 3. Relative efficiencies comparing weighted hotdeck imputation to the adaptive hotdeck imputation

Sample Size	α	β	σ^2	$\hat{V}$	$\hat{V}$ from paper (Reilly1993)	Relative Efficiency
100	0	0.000	0.25	0.275	0.166	0.604
100	1	0.000	0.25	0.310	0.165	0.532
100	0	1.000	0.25	0.419	0.217	0.518
100	1	1.000	0.25	0.504	0.302	0.599
100	0	0.000	2.00	0.390	0.419	1.074
100	1	0.000	2.00	0.417	0.382	0.916
100	0	1.000	2.00	0.916	0.987	1.078
100	1	1.000	2.00	0.959	0.730	0.761

Table 4. Relative efficiencies comparing hotdeck imputation to the adaptive hotdeck imputation

Sample Size	α	β	σ^2	$\hat{V}$	$\hat{V}$ from paper (Reilly1993)	Relative Efficiency
100	0	0.000	0.25	0.174	0.166	0.954
100	1	0.000	0.25	0.195	0.165	0.846
100	0	1.000	0.25	0.312	0.217	0.696
100	1	1.000	0.25	0.346	0.302	0.873
100	0	0.000	2.00	0.323	0.419	1.297
100	1	0.000	2.00	0.309	0.382	1.236
100	0	1.000	2.00	0.823	0.987	1.199
100	1	1.000	2.00	0.672	0.730	1.086

Table 5. Relative efficiencies comparing weighted hotdeck imputation to the Pepe-Flemming estimator

Sample Size	α	β	σ^2	$\hat{V}$	$\hat{V}$ from paper (Pepe and Fleming1991)	Relative Efficiency
200	0	0.000	0.25	0.113	0.033	0.292
200	0	0.693	0.25	0.192	0.051	0.266
200	0	0.000	1.00	0.133	0.053	0.398
200	0	0.693	1.00	0.225	0.088	0.391
100	0	0.000	0.25	0.766	0.072	0.094
100	0	0.693	0.25	1.084	0.118	0.109
100	-1	0.000	0.25	3.447	0.099	0.029
100	-1	0.693	0.25	2.782	0.225	0.081
100	0	0.000	1.00	0.712	0.087	0.122
100	0	0.693	1.00	22.137	0.211	0.010
100	-1	0.000	1.00	1.277	0.145	0.114
100	-1	0.693	1.00	1.834	0.240	0.131

Table 6. Relative efficiencies comparing hotdeck imputation to the Pepe-Flemming estimator

Sample Size	α	β	σ^2	$\hat{V}$	$\hat{V}$ from paper (Pepe and Fleming1991)	Relative Efficiency
200	0	0.000	0.25	0.077	0.033	0.429
200	0	0.693	0.25	0.112	0.051	0.455
200	0	0.000	1.00	0.113	0.053	0.469
200	0	0.693	1.00	0.173	0.088	0.509
100	0	0.000	0.25	0.273	0.072	0.264
100	0	0.693	0.25	0.570	0.118	0.207
100	-1	0.000	0.25	4.147	0.099	0.024
100	-1	0.693	0.25	1.159	0.225	0.194
100	0	0.000	1.00	0.395	0.087	0.220
100	0	0.693	1.00	20.427	0.211	0.010
100	-1	0.000	1.00	0.602	0.145	0.241
100	-1	0.693	1.00	1.497	0.240	0.160

Table 7. Prevalence estimates using Multiple Imputation

		Prevalence Estimates	
		π_f (SE)	π_m (SE)
Hotdeck	2 strata	0.3645 (0.053)	0.2268 (0.065)
	4 strata	0.3687 (0.056)	0.2365 (0.057)
	8 strata	0.3563 (0.055)	0.2234 (0.053)
Weighted	2 strata	0.3651 (0.055)	0.2229 (0.065)
Hotdeck	4 strata	0.3769 (0.056)	0.2198 (0.063)
	8 strata	0.3783 (0.049)	0.2229 (0.067)