Made available by Hasselt University Library in https://documentserver.uhasselt.be

Local multiple imputation

Peer-reviewed author version

AERTS, Marc; CLAESKENS, Gerda; HENS, Niel & MOLENBERGHS, Geert (2002)
Local multiple imputation. In: Biometrika, 89(2). p. 375-388.

DOI: 10.1093/biomet/89.2.375
Handle: http://hdl.handle.net/1942/271

Local Multiple Imputation

By MARC AERTS

Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium

marc.aerts@luc.ac.be

GERDA CLAESKENS

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.

gerda@stat.tamu.edu

NIEL HENS

Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium

niel.hens@luc.ac.be

and GEERT MOLENBERGHS

Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium

geert.molenberghs@luc.ac.be

Summary

Dealing with missing data via parametric multiple imputation methods usually implies stating several strong assumptions about both the distribution of the data and about underlying regression relationships. If such parametric assumptions do not hold, the multiply imputed data are not appropriate and might produce inconsistent estimators and thus misleading results. In this paper, a fully nonparametric and a semiparametric imputation method are studied, both based on local resampling principles. It is shown that the final estimator, based on these local imputations, is consistent under no or fewer parametric assumptions. Asymptotic expressions for bias, variance and mean squared error are derived, showing the theoretical impact of the different smoothing parameters. Simulations illustrate the usefulness and applicability of the method.

Some key words: Bootstrap; Kernel weights; Multiple imputation; Missing value; Nonparametric imputation; Nonresponse; Semiparametric imputation.

1 Introduction

There exist many ways to deal with missing data problems, ranging from the most naive one of focusing on the complete cases only to well-defined parametric, semiparametric and nonparametric approaches. Our approach is novel in several aspects. We focus attention on nonparametric smoothing methods to obtain multiple imputation estimators in a non-Bayesian framework. Unlike most of the literature, which deals with missing covariate values, our method allows for missing response data.

From the large literature on missing data, we highlight only a few particularly relevant references. Kernel methods for imputation of missing values were introduced by Titterington & Sedransk (1989), who used kernel density estimation in combination with a nonparametric bootstrap for imputing values. Their method does not directly account for relationships between variables. For single imputation in a nonresponse setting, Cheng (1994) and Chu & Cheng (1995) used kernel estimators in a regression model, providing a nonparametric version of the so-called 'poor man's data augmentation', which is known to underestimate variability, especially in cases with substantial missingness. For missing covariate data, smoothing methods have been applied by Wang et al. (1998) to estimate selection probabilities. Other semiparametric approaches, in the sense of not having to specify a fully parametric model, although not directly in a smoothing context, are constructed for drop-out models in Scharfstein et al. (1999).

We focus on scenarios where some of the variables are fully observed and some involve missing measurements. The parameter of interest is a marginal parameter of an incompletely observed variable, and the regression relationship between a partially observed response variable and a fully observed covariate is exploited to augment the data. Whereas multiple imputation is mainly regarded as a Bayesian technique (Rubin, 1978, 1987), the proposed methods are bootstrap-based (Efron, 1994). In the next section, we introduce two classes of local bootstrap methods. The local resampling method is fully nonparametric and hence relaxes distributional assumptions and assumptions concerning regression functions. The local semiparametric resampling method still assumes that the conditional distributions are, for example, locally normal but allows nonlinear conditional mean structures. Throughout, we assume an ignorable nonresponse mechanism.

In the next section we introduce some basic notation and explain the imputation algorithm. Section 3 focuses on the local bootstrap method, asymptotic results are presented

in § 4 and the selection of the different smoothing parameters involved in the procedure, is considered in § 5. Section 6 summarises the results of a simulation study.

2 Local imputation scheme

Consider one completely observed continuous variable X and one incompletely observed continuous variable Y. The parameter of interest is a function $\theta(\mu_X, \mu_Y)$ of the two means. Since there is no missing X value, the problem reduces to consistent estimation of $\mu = \mu_Y$; estimators of other moments of Y and functions thereof can be obtained in a straightforward manner. The main idea is to exploit the assumed regression relationship between X and Y to yield better estimators for μ. Let $Z_i = (X_i, Y_i, \delta_i), i = 1, \ldots, n$, be independent observations, where $\delta_i = 0$ if Y_i is missing and $\delta_i = 1$ otherwise. Under the strongly ignorable missing at random assumption (Rosenbaum & Rubin, 1983)

$$\pi(X) := E(\delta|X) = E(\delta|X, Y). \tag{1}$$

In other words, Y and δ are conditionally independent given X. Note that this assumption is weaker than missingness completely at random since dependence on the observed variable X is allowed. Little & Rubin (1987, p. 15), term data of this type missing at random but not observed at random.

Our approach extends the local single imputation of Cheng (1994) to a non- or semi-parametric version of a 'proper' imputation method and is related to the approximate Bayesian bootstrap method as described in equation (3.7) of Efron (1994); see also Little & Rubin (1987, § 12.4). An essential ingredient of the algorithm is the local generation of Y observations. Let x be a specific value of X at which a Y value is to be generated and let $w_j(x), j = 1, \ldots, n$, denote positive weights with $\sum_{j=1}^{n} w_j(x) = 1$. The local resampling method generates a Y value from the distribution $\mathcal{L}(x)$ with cumulative distribution function

$$\sum_{j=1}^{n} w_j(x) I\{Y_j \leq y\}. \tag{2}$$

Detailed treatment of the choice of weights is postponed to § 3. First we describe the steps of the local m-fold multiple imputation algorithm, where as an example, attention is restricted to a normal likelihood in Step 2.

Step 1: Resampling step

Fix ℓ between 1 and m. For each observation $i = 1, \ldots, n$, if $\delta_i = 1$, generate $Y_i^*(\ell)$ from the distribution $\mathcal{L}(X_i)$. This is a nonparametric resampling of the observed data vectors.

Step 2: Imputation step

Fix ℓ between 1 and m. Given the data from Step 1, we create imputations for the missing Y values. This can be done in several ways, using local resampling or local semiparametric resampling. More explicitly, conditional on the resampled data $(X_i, Y_i^*(\ell), \delta_i), i = 1, \ldots, n$, we construct a distribution $\mathcal{L}_\ell^*(X_i)$, for local resampling, or local estimators $\hat{\mu}_\ell^*(X_i)$, $\hat{\sigma}_\ell^{*2}(X_i)$, for local semiparametric resampling. If Y_i is missing, that is if $\delta_i = 0$, we generate $Y_i^+(\ell)$ from $\mathcal{L}_\ell^*(X_i)$, for local resampling, or, for local semiparametric resampling, we generate $Y_i^+(\ell)$ from $N\{\hat{\mu}_\ell^*(X_i), \hat{\sigma}_\ell^{*2}(X_i)\}$. It is clear that local semiparametric resampling is more efficient if normality holds. In both Step 1 and Step 2, data are generated independently for $i = 1, \ldots, n$, $\ell = 1, \ldots, m$.

Step 3: Construction of the final estimators

For $\tilde{Y}_i(\ell) = \delta_i Y_i + (1 - \delta_i) Y_i^+(\ell)$, $\hat{\mu}(\ell) = n^{-1} \sum_{i=1}^n \tilde{Y}_i(\ell)$ is the estimator of the mean based on the ℓth augmented dataset, and the final multiple-imputation estimator for μ is

$$\hat{\mu} = \frac{1}{m} \sum_{\ell=1}^{m} \hat{\mu}(\ell).$$

The algorithm has the same structure as its parametric counterpart. Since an imputed observation Y^+ is subject to extra variability, Step 1 is needed for obtaining a proper imputation method (Efron, 1994). This extra randomness can be introduced in different ways. The triplets $(X_i, Y_i, \delta_i), i = 1, \ldots, n$, could be resampled with replacement; this is case resampling. We opted for an alternative approach, where Y values are generated, conditional on X and δ, incorporating the regression relationship between X and Y in a nonparametric way. This approach is legitimate because of assumption (1), which states that the missingness mechanism is noninformative for the parameter μ of interest. The advantage of the method is that it generates samples with exactly the same range of X values as in the original sample, avoiding samples which might only poorly reflect the regression structure, the latter which is essential in Step 2.

The nonparametric imputation method is applicable in a wide variety of statistical models, and can be used for discrete response data. The small adaptation needed for semiparametric resampling is the specification of the appropriate distribution function in

Step 2 of the algorithm. For examples of local likelihood estimators in multi-parameter families, see Aerts & Claeskens (1997).

A local bootstrap method both avoids parametric assumptions and allows much more flexibility in the regression design than does for example hot-deck imputation (Rao & Shao, 1992), where one requires a covariate to take on only a few different values, with replication.

3 Local bootstrap methods

The choice of the weights in the resampling scheme is crucial. Global uniform weights $\delta_j / \sum_j \delta_j$ would simply result in mean imputation of the Y-values, ignoring the regression structure completely. More useful are kernel weights of the type

$$w_j(x) = \frac{\delta_j K_h(x - X_j)}{\sum_{k=1}^{n} \delta_k K_h(x - X_k)} \tag{3}$$

where the kernel $K(\cdot)$ is a symmetric unimodal probability density function, $K_h(u) = K(u/h)/h$, and $h = h_n$ is a bandwidth parameter converging to zero as the sample size increases. It is not necessary to use the same set of weights in the resampling and imputation steps. In particular, since the smoothing weights in Step 2 use a resampled set of data, it is advisable to use a second bandwidth $g = g_n$ in a possibly different kernel L for the construction of the weights in the imputation step. In case of possible confusion, choice of bandwidth will be included in the notation.

Local weights (3) are defined such that observed Y_j values, of which the corresponding X_j is closer to the specific value x, and which are in an area with larger chance of having missing observations, get larger weights. The latter is readily understood by rewriting the weights (3) as $w_j(x) = \delta_j \tilde{w}_j(x)/\hat{\pi}(x)$ where the classical Nadaraya-Watson weights $\tilde{w}_j(x) = K_h(x - X_j) / \sum_{k=1}^{n} K_h(x - X_k)$ and where $\hat{\pi}(x) = \sum_{j=1}^{n} K_h(x - X_j)\, \delta_j / \sum_{j=1}^{n} K_h(x - X_j)$ is the kernel estimator for $\pi(x)$. Thus we do not have to make any parametric assumptions about the missingness probability distribution since this is automatically taken care of by the kernel weights. The effect of $\hat{\pi}(x)$ on the weights stresses the importance of the few available but highly informative Y observations in a 'sparse' area with a lot of missingness.

In the complete data case, Aerts et al. (1994) have shown that distribution (2) is consistent and asymptotically normal for estimating the conditional distribution of Y given $X = x$. They also showed that a resampling scheme based on this distribution leads to a consistent bootstrap procedure. In an analogous way it can be shown that, if Y values are

missing, (2) is a consistent estimator for the distribution function $P(Y \leq y|X = x, \delta = 1)$. Its mean equals the well-known Nadaraya-Watson estimator at x from the complete cases (Nadaraya, 1964; Watson, 1964), which can be rewritten as $\hat{\mu}(x) = \sum_{j=1}^{n} \tilde{w}_j(x)\ \delta_j Y_j / \hat{\pi}(x)$, where the numerator is the kernel estimator of $E(\delta Y|X = x)$ and the denominator estimates $E(\delta|X = x)$ nonparametrically. It immediately follows from assumption (1), that $\hat{\mu}(x)$ is an estimator of $E(Y|X = x)$. The variance of (2) is a consistent nonparametric variance estimator. These provide alternatives to the local likelihood estimators in local semiparametric resampling.

Given the known limitations of Nadaraya-Watson weights, alternative sets of local weights are worth considering, such as biased bootstrap weights (Hall & Presnell, 1999), constrained to make the adjusted estimator unbiased for linear functions. Here we define $\breve{w}_j(x) = \delta_j K_h(x - X_j) \left\{1 + c(x - X_j)K_h(x - X_j)\right\}^{-1} \left[\sum_{k=1}^{n} \delta_k K_h(x - X_k) \left\{1 + c(x - X_k)\right.\right.$ $\left.\left. \times\ K_h(x - X_k)\right\}^{-1}\right]^{-1}$, where c is the solution to the equation $\sum_{j=1}^{n} \delta_j (x - X_j) K_h(x - X_j) \left\{1 + c(x - X_j)K_h(x - X_j)\right\}^{-1} = 0$. These weights are asymptotically equivalent to local linear weights. Hence they automatically correct for boundary bias, while remaining positive. Alternative or additional constraints on the resampling distribution can be imposed in a similar way.

If the proportion of missingness would be known, missing data could be dealt with as in a weighted-distributions regression setting, for which Ahmad (1995), see also Jones (1991), derives a kernel estimator analogue to the direct sampling case. The corresponding weights, with π estimated by the kernel estimator $\hat{\pi}$, are defined as $\breve{w}_j(x) = \delta_j \tilde{w}_j(x) \{\hat{\pi}(X_j) \sum_{k=1}^{n} \delta_k \tilde{w}_k(x) / \hat{\pi}(X_k)\}^{-1}$. The important difference from the weights (3) is the evaluation of $\hat{\pi}$ at the covariates X_j.

The performance of the above weights will be numerically illustrated in § 6.1, where it turns out that the precise choice of weights has little effect on the final estimator.

If more than one variable is completely observed, local methods could take all of them into account. However, in high dimensions kernel-based methods might lose some of their attractiveness because of the curse of dimensionality.

4 Asymptotic expressions of bias and variance

The final estimator $\hat{\mu}$ is consistent, under conditions similar to those in Cheng (1994). Smoothness conditions require $\mu(x)$, the conditional mean of Y given $X = x$, the den-

sity function, $f_X(x)$, and the function $\pi(x)$, to possess at least two bounded derivatives, bandwidth sequences to tend to zero at a rate faster than $n^{-1/3}$, and kernel functions K and L in both steps to be bounded and symmetric probability density functions with finite second moments. We also assume that Y has a finite second moment, and that all required expected values are finite.

A first result shows that the final estimator $\hat{\mu}$ is asymptotically unbiased and that the bias depends on both bandwidth sequences in a typical nonparametric way. For some constants c_1 and c_2,

$$E(\hat{\mu}) = \mu + c_1 h^2 + c_2 g^2 + o(h^2 + g^2) \ , \quad \text{as } n \to \infty. \tag{4}$$

For the asymptotic variance of $\hat{\mu}$, we get as $n \to \infty$, with additional constants $c_3, \ldots, c_6$ and with $\sigma^2(X) = \text{var}(Y|X)$,

$$\begin{aligned} \text{var}(\hat{\mu}) \ &= \ (mn)^{-1} E[\sigma^2(X)\{1-\pi(X)\}/\pi(X)] + n^{-1}[E\{\sigma^2(X)/\pi(X)\} + \text{var}\{\mu(X)\}] \\ &\quad + n^{-2}(c_3 h^{-1} + c_4 g^{-1}) + n^{-1}(c_5 h^2 + c_6 g^2) + o\{(h^2+g^2)n^{-1}\}, \end{aligned} \tag{5}$$

showing that $\hat{\mu}$ is root-n consistent as an estimator for μ. Outlines of proofs of (4) and (5) are given in the Appendix. The second term on the right-hand side of $\text{var}(\hat{\mu})$ represents the variance of a single mean imputation, as shown in Cheng (1994). The first term stems from the multiple imputation approach with additional randomness generated in Step 1. In the case of no missingness, the leading term in (5) reduces to $\text{var}(Y)/n$, as expected. The constants c_i depend on the second derivatives of $\mu(x)$, $f_X(x)$ and $\pi(x)$, with respect to x, as well as on second moments of the kernel functions.

The following central limit result holds

$$\sqrt{n}\,\text{var}(\hat{\mu})^{-1/2}\{\hat{\mu} - E(\hat{\mu})\} \to N(0,1), \tag{6}$$

in distribution, with mean and variance as given by (4) and (5); see the Appendix for more details.

The mean squared error of $\hat{\mu}$ is

$$\text{MSE}(\hat{\mu}) = c_0 n^{-1} + (c_1 h^2 + c_2 g^2)^2 + (c_3 h^{-1} + c_4 g^{-1})n^{-2} + (c_5 h^2 + c_6 g^2)n^{-1} + o\{(h^2+g^2)n^{-1}\}, \tag{7}$$

where $c_0 = m^{-1}E[\sigma^2(X)\{1-\pi(X)\}/\pi(X)] + E\{\sigma^2(X)/\pi(X)\} + \text{var}\{\mu(X)\}$.

In the remainder of this section we examine the behaviour of a particular estimator of $\mathrm{var}(\hat{\mu})$ by showing how it relates to expression (5).

In parametric multiple imputation estimation, the variance of $\hat{\mu}$ is typically estimated by $S^2(\hat{\mu}) = \hat{W} + (1+m^{-1})\hat{B}$, where $\hat{W}$ is the average within-imputation variance estimator, i.e. $\hat{W} = m^{-1}\sum_{\ell=1}^{m} S_\ell^2$, where nS_ℓ^2 is the unbiased sample variance within the ℓth augmented dataset, and $\hat{B}$ is the between-imputation variance, i.e. $\hat{B} = (m-1)^{-1}\sum_{\ell=1}^{m}\{\hat{\mu}(\ell) - m^{-1}\sum_{k=1}^{m}\hat{\mu}(k)\}^2$. It is shown in the Appendix that

$$E(\hat{W}) = \frac{1}{n}\left[\mathrm{var}\{\mu(X)\} + E\{\sigma^2(X)\}\right] + O\{(h^2+g^2)n^{-1}\} \quad (8)$$

$$E(\hat{B}) = \frac{1}{n}E\left\{\frac{1-\pi(X)}{\pi(X)}\sigma^2(X)\right\} + O\{(h^2+g^2)n^{-1}\}, \quad (9)$$

which proves the asymptotic unbiasedness of $S^2(\hat{\mu})$ as an estimator of $\mathrm{var}(\hat{\mu})$.

The construction of the variance estimator $S^2(\hat{\mu})$ is simple and is exactly the same as in parametric multiple imputation methods. This is an advantage over other estimators of this variance, such as the nonparametric estimator of Cheng (1994), where an additional smoothing parameter needs to be selected.

5 Optimal bandwidths

5.1 Asymptotically optimal bandwidths

The asymptotically optimal bandwidths minimise the dominant terms in (7). Terms of order $O(h/n)$ are negligible compared to order $O\{(nh)^{-2}\}$ terms, as long as $h = O(n^{-\alpha})$, with $\alpha > 1/3$. The same holds for g, where the order of g is not restricted to be the same as the order of h.

By differentiating (7) and omitting all negligible terms, we find that both bandwidths are $O(n^{-2/5})$, yet with different constants, depending on c_1, c_2, c_4 and c_5.

Since the constants in front of the $n^{-2/5}$ are functions of higher derivatives of $\mu(x)$, these cannot be computed exactly for any dataset. Data-driven bandwidth selection is to be advised, although in practice any 'reasonable' bandwidth choice will give satisfactory results.

5.2 Jackknife bandwidth selection

Using the asymptotically optimal order derived in the previous section, jackknife ideas can be utilized to estimate the mean squared error of $\hat{\mu}$ for different choices of the bandwidths h and g. A data driven selection of both smoothing parameters minimizes the estimated mean squared error.

As a result of the double resampling estimation, the following jackknife procedure retains all data generated in Steps 1 and 2, but modifies, for each $i = 1, \ldots, n$, the imputed data $Y_j^+(\ell)$ to $Y_j^{+(-i)}(\ell)$ by shifting them to a new mean reflecting the deletion of the ith observation while using $h_{n-1} = C_h(n-1)^{-2/5}$ and $g_{n-1} = C_g(n-1)^{-2/5}$. This idea was inspired by the adjusted jackknife as proposed by Rao & Shao (1992). Here and in the sequel, a superscript $(-i)$ refers to exclusion of the ith observation.

Based on all the data, the conditional mean of the variable $Y_i^+(\ell)$ is given by $\hat{\mu}_n^+(X_i; h_n, g_n) = \sum_{k=1}^n w_k(X_i; g_n)\hat{\mu}(X_k; h_n)$ where $\hat{\mu}(X_k; h_n) = \sum_{j=1}^n w_j(X_k; h_n)Y_j$. Within each jackknife run i, referring to deletion of the ith observation, the imputed observation $Y_j^+(\ell)$ is replaced by the adjusted imputed value

$$Y_j^{+(-i)}(\ell) = Y_j^+(\ell) + \{\hat{\mu}_{n-1}^{+(-i)}(X_j; h_{n-1}, g_{n-1}) - \hat{\mu}_n^+(X_j; h_n, g_n)\},$$

where the notation explicitly mentions the bandwidths. Estimator $\hat{\mu}^{(-i)}$ is $\hat{\mu}$ based on jackknife imputed values. The leading bias term of the average $\bar{\mu}$ of the jackknife pseudovalues, $\hat{\mu}_{n,i} = \{n^\alpha\hat{\mu} - (n-1)^\alpha\hat{\mu}^{(-i)}\}/\{n^\alpha - (n-1)^\alpha\}$, cancels out when $\alpha = 4/5$, and leads to a bias-corrected estimator $\bar{\mu}$, which is called the generalised jackknife statistic (Gray & Schucany, 1972). Since the bias of $\hat{\mu}$ is not $O(1/n)$, the choice $\alpha = 1$ corresponding to Quenouille's (1956) original jackknife pseudovalues is not appropriate here.

The difference $\hat{b}(\hat{\mu}) = \hat{\mu} - \bar{\mu}$ is known as the jackknife bias estimator and the jackknife variance estimator for $\hat{\mu}$ is given by $\hat{var}(\hat{\mu}) = \{n(n-1)\}^{-1}\sum_{i=1}^n(\hat{\mu}_{n,i} - \bar{\mu})^2$.; see Efron & Tibshirani (1993, § 11.2). Both bias and variance depend on the values of the unknown constants C_h and C_g. Optimal choices can now be derived by minimising the estimated mean squared error, given by $\hat{b}^2(\hat{\mu}) + \hat{var}(\hat{\mu})$. As illustrated in the next section, this jackknife procedure succeeds in selecting a proper choice of C_h and C_g. An in-depth study of the theoretical properties and the finite sample behaviour of this jackknife bandwidth selector is beyond the scope of this paper and will be pursued elsewhere.

6 Simulation results

6.1 A simulation study

The following methods for multiple imputation are included in this simulation study. The first, naive approach uses the complete cases only. Among the parametric methods we consider single imputation (Buck, 1960) and multiple imputation, according to Rubin (1978, 1987) and Efron (1994). These methods all assume a parametric regression relationship between Y and X. Rubin's multiple imputation assumes joint normality of (X, Y). In Efron's bootstrap approach, the complete cases are resampled and used to fit a linear regression model of Y on X in order to impute Y-values from a normal distribution with estimated linear conditional mean function and estimated constant variance.

Three nonparametric approaches are also included. The first is a single imputation method, in which a local linear estimator of the conditional mean is used to impute for missing Y values (Cheng, 1994). The other two methods are those studied in this paper, namely multiple imputation by local resampling or local semiparametric resampling, employing different sets of local weights (1) w_j, (2) $\breve{w}_j$, (3) $\breve{w}_j$; see §3.

In a first scenario Y observations are generated from a normal distribution with conditional mean $\mu(x) = -3 + x + 7x^2$ and conditional variance $\sigma^2(x) = \exp(3 + 0.2x)$. The completely observed X variable follows a uniform distribution on the interval [0,10]. Values are missing with conditional probability $1 - \pi(x) = \{1 + \exp(0.5 - 0.1(x - 5)^2)\}^{-1}$, which is largest at the ends of the interval. With these specifications, the true value of the parameter of interest is $\mu = E\{\mu(X)\} = 235.33$ and the total percentage of missingness is $E\{\pi(X)\} = 0.57$. In this and all other scenarios we took the number of multiple imputations to be $m = 3$. Other values, $m = 5$ and $m = 10$, gave very comparable results and are not shown.

We generated 1000 samples $\{(X_i, Y_i, \delta_i), i = 1, \ldots, n\}$. Table 1 summarises the main results for n=200. Results for n=100 are similar and are not shown here. The standard normal kernel function was used in all nonparametric imputation methods and all bandwidths were kept fixed. For the nonparametric single imputation only one bandwidth is needed and was taken as 1.5. For the local semiparametric resampling the bandwidth in Step 1 was $h = 0.25$ and in Step 2 we chose $g = 1.5$. The local resampling method used the same bandwidth $h = g = 0.25$ in both steps. These choices are based on some initial

experiments.

For each imputation method and each run we computed the multiple imputation estimate $\hat{\mu}$, its estimated standard error $S(\hat{\mu})$ and a 95% confidence interval $\hat{\mu} \pm 1.96S(\hat{\mu})$. Averages of the point estimates are shown in columns 1 and 2 of Table 1. Column 3 shows the simulated standard error of $\hat{\mu}$. Columns 4 and 5 show the average length of the 1000 confidence intervals and the simulated coverage probability. Rubin & Schenker (1986) suggested adjusting additionally for the multiple imputation by using critical points based on t with $(m-1)\{1+(m/(m+1))(\hat{W}/\hat{B})\}^2$ degrees of freedom. The average lengths of these adjusted confidence intervals and simulated coverage probabilities are shown in columns 6 and 7, only for the multiple imputation methods.

Table 1 About Here.

Next to linearity of $\mu(x)$, all parametric multiple imputation methods assume a constant variance $\sigma^2(x)$. Moreover Rubin's parametric multiple imputation assumes X to be normally distributed. The local resampling and local semiparametric resampling approaches do not violate any model specifications.

As expected, the complete-cases method and the parametric imputation methods clearly underestimate the true mean μ while the nonparametric approaches perform much better. A comparison of the averages of the estimated standard errors and the simulated standard errors confirms the need for multiple imputation. Note that the average lengths of the confidence intervals and the associated coverage probabilities are equal or larger for the construction based on a t random variable (Rubin & Schenker, 1986) for all multiple imputation methods. This approach reduces to the normal confidence intervals for single imputation.

For this scenario there is not much difference between the local semiparametric resampling and local resampling methods; both improve significantly upon the parametric methods. Also, there are almost no differences between the different local weighting schemes.

In a second scenario, response data follow a 6:4 mixture of $N\{\mu(x), \sigma^2(x)\}$ and $\text{Exp}\{1/\mu(x)\}$, where $\mu(x) = 6 + (x-2)(x-4) + 5\cos(\pi x)$, $\sigma(x) = \exp(0.02x)$, and $\text{logit}\{\pi(x)\}=2-0.4x$, resulting in $\mu = 17.33$. Since there is more misspecification, differences between parametric and nonparametric methods are more pronounced. Table 2 gives the simulation results for the local methods using bandwidths $h = 1$ and $g = 1.5$ for $n = 200$.

TABLE 2 ABOUT HERE.

In this scenario, the semiparametric methods, using a normal local likelihood in Step 2, turn out to be quite robust against the model misspecification. The local linearised weights $\breve{w}_j$ result in somewhat lower coverage probabilities, caused by a slight overestimation of μ. Better results might be obtained if the bandwidth were be optimised in each simulation run. The precise choice of local weights turns out to be of less importance.

Several parameters and underlying functions may influence the behaviour of the different imputation methods. We experimented with some other variations of these simulation settings, all leading to essentially the same conclusions. The local imputation method improves upon the classical methods when one or more of the parametric assumptions are violated. When all assumptions underlying the parametric multiple imputation methods are fulfilled, local resampling and local semiparametric resampling do not outperform the parametric approaches, although the loss in efficiency incurred by using unnecessarily a local imputation method remains small.

6.2 Jackknife data driven bandwidth selection

As an illustration, we applied the jackknife method of § 5.2 to a randomly chosen sample obtained from the first scenario in § 6.1, using the weights w_j, defined in (3). For the local resampling imputation, the grid 0.2, 0.25, 0.3, 0.5, 1, 2.5, 5, 20, 30, 40 was used for both constants C_h and C_g. In this way 100 estimates of $\hat{\mu}$ and the corresponding mean squared error of $\hat{\mu}$ were calculated. This resulted in a surface as shown in Fig.1(a). Figure 1(b) shows the estimated mean squared error as a function of $\hat{\mu}$ using a loess fit. This shows that lower values of the mean squared error correspond to estimates in the neighbourhood of the true value $\mu = 235.33$. The minimum is attained at $\hat{\mu} = 233.15$, with bandwidths $h = 0.601$ and $g = 0.024$. This latter plot also shows that different choices for h and g can lead to a wide range of $\hat{\mu}$-values, from about 225 to 260, indicating that a precise bandwidth choice is not unimportant.

Jackknifing with local semiparametric imputation was also examined for the same sample, using a C_h, C_g-grid based on 1, 1.5, 2.5, 5, 7.5, 10, 15, 20, 30, 40. Larger values were needed, which seems plausible for a partly parametric approach. A plot of the estimated mean squared error versus $\hat{\mu}$ is shown in Fig.1(c). The loess curve indicates a steeper descent towards the minimum, but on the other hand there is more variability. Estimates in

the range of 224 to 231 have more or less the same associated mean squared error. For this sample, however, the local resampling method seems to do better. A similar experiment was done with sample size equal to 100 instead of 200. The result is shown in Fig.1(d). The curve seems to flatten out at its minimum of $\hat{\mu} = 236.39$, corresponding to $h = 0.396$ and $g = 3.170$.

Our conclusion is that the jackknife method gives promising results but further research is needed.

FIGURE 1 ABOUT HERE.

APPENDIX

Proofs

Throughout the Appendix, we denote by $E(\cdot|O)$ the expectation conditional on $Z_1, \ldots, Z_n$ and by $E(\cdot|O, R)$ the expectation conditional on $Z_1, \ldots, Z_n, Z_1^*, \ldots, Z_n^*$, where $Z_i^* = (X_i, Y_i^*, \delta_i)$.

Proofs are given here for the local resampling algorithm; for local semiparametric resampling, the proofs are very similar. Arguments hold for any set of weights on the observed data for which, with g a twice continuously differentiable function, $\sum_{j=1}^n E\{w_j(X;h)g(X_j)\} \to E\{g(X)\} + O(h^2)$. This condition holds for the weights studied in § 3.

Proof of (4)

Since $\hat{\mu}$ is defined as

$$\hat{\mu} = \frac{1}{m}\sum_{\ell=1}^{m}\frac{1}{n}\sum_{j=1}^{n}\tilde{Y}_j(\ell),$$

where $\tilde{Y}_j(\ell) = \delta_j Y_j + (1-\delta_j)Y_j^+(\ell)$, the first term contributes to $E(\delta_j Y_j) = E\{E(\delta_j Y_j|X_j)\} = E\{\pi(X)\mu(X)\}$. Next, by definition of $Y_j^+(\ell)$, $E\{(1-\delta_j)Y_j^+(\ell)\} = E[(1-\delta_j)E\{Y_j^+(\ell)|O,R\}] = E\{(1-\delta_j)\hat{\mu}_\ell^*(X_j;g)\}$. Using the explicit formula $\hat{\mu}_\ell^*(X_j;g) = \sum_{i=1}^n w_i(X_j;g)Y_i^*(\ell)$, conditioning on the observed data and using a Taylor expansion, hereby making use of the symmetry of the kernel functions K and L, we obtain that

$$E\{(1-\delta_j)\tilde{Y}_j(\ell)\} = E\{\mu(X)(1-\pi(X))\} + O(h^2+g^2).$$

Together with the result for the first term, $E(\delta_j Y_j)$, this concludes the proof.

Proof of (5)

Conditioning on observed and first-stage resampled data, we obtain $\text{var}(\hat{\mu}) = E\{\text{var}(\hat{\mu}|O,R)\} + \text{var}\{E(\hat{\mu}|O,R)\}$. By definition of the multiple imputation estimator $\hat{\mu}$, we have

$$\begin{aligned} E\{\text{var}(\hat{\mu}|O,R)\} &= \frac{1}{mn} E[(1-\delta_1)\,\text{var}\{Y_1^+(1)|O,R\}] \\ &= \frac{1}{mn} E[(1-\delta_1) E\{\hat{\sigma}_1^{*2}(X_1;g)|O\}], \end{aligned} \tag{A}$$

where

$$\hat{\sigma}_1^{*2}(X_1;g) = \sum_{j=1}^{n} \{Y_j^*(1) - \hat{\mu}_1^*(X_1;g)\}^2 w_j(X_1;g).$$

The inner expectation in (A), which is conditional on the observed data, is most easily obtained by explicitly rewriting $\{Y_j^* - \hat{\mu}^*(X_i;g)\}^2$ as $(Y_j^*)^2 - 2Y_j^*\hat{\mu}^*(X_i;g) + \{\hat{\mu}^*(X_i;g)\}^2$, and by calculating the conditional expectation of each term separately, using computations similar to those in the proof of (4). Proceeding this way, we obtain that

$$E\{\text{var}(\hat{\mu}|O,R)\} = \frac{1}{mn} E[\{1-\pi(X)\}\sigma^2(X)] + O\{(h^2+g^2)n^{-1}\}.$$

Next, we turn to

$$\begin{aligned} &\text{var}\{E(\hat{\mu}|O,R)\} = \text{var}[E\{E(\hat{\mu}|O,R)|O\}] + E[\text{var}\{E(\hat{\mu}|O,R)|O\}] \\ &= \tfrac{1}{n}\,\text{var}[\delta_1 Y_1 + (1-\delta_1) E\{\hat{\mu}_1^*(X_1;g)|O\}] + \tfrac{1}{mn} E[(1-\delta_1)\,\text{var}\{\hat{\mu}_1^*(X_1;g)|O\}]. \end{aligned} \tag{B}$$

Similar calculations as before yield, for the first term in (B),

$$\tfrac{1}{n}[E\{\sigma^2(X)/\pi(X)\} + \text{var}\{\mu(X)\} + O(h^2+g^2)],$$

and, for the second term,

$$\tfrac{1}{mn}\left(E[\{1-\pi(X)\}^2\sigma^2(X)/\pi(X)] + O(h^2+g^2)\right),$$

from which the result follows.

Proof of (6)

Define

$$V_{1n} = \tfrac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\hat{\mu}_1^*(X_i;g) - E\{\mu(X)\} + \tfrac{1}{n}\sum_{i=1}^{n}\delta_i Y_i,$$

$$V_{2n} = E(V_{1n}|O).$$

Conditional on $Z_1^*, \ldots, Z_n^*$, $\sqrt{n}(\hat{\mu} - \mu - V_{1n}) \to N_1$, in distribution and, conditional on $Z_1, \ldots, Z_n$, $\sqrt{n}(V_{1n} - V_{2n}) \to N_2$, in distribution. Unconditionally, $\sqrt{n}V_{2n} \to N_3$, in distribution, where the N_i have a normal distribution. Since V_{2n} features only observed data, normality is readily obtained. Distributions of N_1 and N_2 can be obtained by separating the randomness induced by the bootstrap resampling as in the following triangular arrays:

$$\begin{aligned}
\sqrt{n}(V_{1n} - V_{2n}) &= \sum_{j=1}^{n} \Bigg[n^{-1/2}(1-\delta_j) \sum_{i=1}^{n} w_i(X_j; g) \sum_{\ell=1}^{n} \Big\{ Y_\ell - \sum_{k=1}^{n} w_k(X_i; h) Y_k \Big\} \\
&\qquad \times I\{ w_{(\ell-1)}(X_j; h) < U_{1j} \le w_{(\ell)}(X_j; h) \} \Big], \\
\sqrt{n}(\hat{\mu} - \mu - V_{1n}) &= \sum_{i=1}^{n} \Bigg[n^{-1/2}(1-\delta_i) \sum_{\ell=1}^{n} \{ Y_\ell^* - \sum_{k=1}^{n} w_k(X_i; g) Y_k^* \} \\
&\qquad \times I\{ w_{(\ell-1)}(X_i; g) < U_{2\ell} \le w_{(\ell)}(X_i; g) \} \Big],
\end{aligned}$$

where $w_{(k)}(X_i; h) = \sum_{j=1}^{k} w_j(X_i; h)$ and $w_{(0)} = 0$. The independent random variables U_{1j} (respectively U_{2j}) follow a uniform distribution on (0, 1), and are independent of $Z_1, \ldots, Z_n$ (respectively $Z_1^*, \ldots, Z_n^*$). A central limit theorem result for the triangular arrays above is obtained via classical arguments.

Application twice of Lemma 1 of Schenker & Welsh (1988) yields the desired result that $\sqrt{n}(\hat{\mu} - \mu) - N \to 0$ in distribution, where N is as the convolution of the three distributions above, namely a normal random variable with mean and variance as already calculated in (4) and (5).

Proof of (8)

By straightforward calculation we get that, with $\mu_2(X) = E(Y^2|X)$,

$$\begin{aligned}
E(\hat{W}) &= \frac{1}{n(n-1)} \sum_{i=1}^{n} E \left[\{\delta_i Y_i + (1-\delta_i) Y_i^+(1)\}^2 - \frac{1}{n} \left\{ \sum_{i=1}^{n} \delta_i Y_i + (1-\delta_i) Y_i^+(1) \right\}^2 \right] \\
&= \tfrac{1}{n} E\{\mu_2(X)\} - \tfrac{1}{n} \left[E\{\pi(X)\mu(X)\} \right]^2 - \tfrac{2}{n} E\{\pi(X)\mu(X)\} \cdot E[\{1-\pi(X)\}\mu(X)] \\
&\quad - \tfrac{1}{n} (E[\{1-\pi(X)\}\mu(X)])^2 + O\{(h^2+g^2)n^{-1}\} \\
&= \tfrac{1}{n} E\{\mu_2(X)\} - \tfrac{1}{n} [E\{\mu(X)\}]^2 + O\{(h^2+g^2)n^{-1}\} \\
&= \tfrac{1}{n} [\mathrm{var}\{\mu(X)\} + E\{\sigma^2(X)\}] + O\{(h^2+g^2)n^{-1}\},
\end{aligned}$$

which is result (8).

Proof of (9)

For $\ell = 1, \ldots, m$, define the random variables

$$D_n(\ell) = \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)Y_i^+(\ell).$$

Being a sample variance, the estimator $\hat{B}$ is an unbiased estimator of the variance of $D_n(1)$, conditional on the observed data. Hence,

$$E(\hat{B}) = E[\mathrm{var}\{D_n(1)|O,R\}] + E(\mathrm{var}[E\{D_n(1)|O,R\}|O]). \qquad \text{(C)}$$

By definition of $D_n(1)$ and $Y_i^+(1)$,

$$\mathrm{var}\{D_n(1)|O,R\} = \frac{1}{n^2}\sum_{i=1}^{n}(1-\delta_i)\,\mathrm{var}\{Y_i^+(1)|O,R\} = \frac{1}{n^2}\sum_{i=1}^{n}(1-\delta_i)\hat{\sigma}_1^{*2}(X_i;g).$$

Since this depends on both the observed and the first-stage resampled data, we calculate the first term of (C) via $E(E[\mathrm{var}\{D_n(1)|O,R\}|O])$. As in the proof of (5), the expectation of the resulting random variable is given by

$$\tfrac{1}{n}E[\{1-\pi(X)\}\sigma^2(X)] + O\{(h^2+g^2)n^{-1}\}.$$

The second term in (C) can be shown to equal

$$\tfrac{1}{n}E[\{1-\pi(X)\}^2\sigma^2(X)/\pi(X)] + O\{(h^2+g^2)n^{-1}\},$$

from which (9) follows.

Acknowledgement

We gratefully acknowledge support from the Fund for Scientific Research - Flanders (Belgium). We also wish to thank the referees and editors for their helpful comments.

References

Aerts, M. & Claeskens, G. (1997). Local polynomial estimation in multiparameter likelihood models. *J. Am. Statist. Assoc.* **92**, 1536–45.

Aerts, M., Janssen, P. & Veraverbeke, N. (1994). Bootstrapping regression quantiles. *J. Nonparam. Statist.* **4**, 1–20.

Ahmad, I.A. (1995). On multivariate kernel estimation for samples from weighted distributions. *Statist. Prob. Lett.* **22**, 121–9.

Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J. R. Statist. Soc.*B, **22**, 302–6.

Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random. *J. Am. Statist. Assoc.* **89**, 81–7.

Chu, C.K. & Cheng, P.E. (1995). Nonparametric regression estimation with missing data. *J. Statist. Plan. Infer.* **48**, 85–99.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *J. Am. Statist. Assoc.* **89**, 463–75.

Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Gray, H.L. & Schucany, W.R. (1972). *The Generalized Jackknife Statistic.* Statistics Textbooks and Monographs, Vol. 1. New York: Marcel Dekker.

Hall, P. & Presnell, B. (1999). Intentionally biased bootstrap methods. *J. R. Statist. Soc.*B, **61**, 143–58.

Jones, M.C. (1991). Kernel density estimation for length biased data. *Biometrika* **78**, 511–9.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York: J. Wiley & Sons.

Nadaraya, E.A. (1964). On estimation regression. *Theory Prob. Applic.* **9**, 141–2.

Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika* **43**, 353–60.

Rao, J.N.K. & Shao, J. (1992) Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79**, 811-22.

Rosenbaum, P.R. & Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **60**, 211–3.

Rubin, D.B. (1978). Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse. In *Imputation and Editing of Faulty or Missing Survey Data*, pp. 1–23. U.S. Department of Commerce.

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

Rubin, D.B. & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J. Am. Statist. Assoc.* **81**, 366-74.

Scharfstein, D.O., Rotnitzky, A. & Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse model (with discussion). *J. Am. Statist. Assoc.* **94**, 1096–146.

Schenker, N. & Welsh, A.H. (1988). Asymptotic results for multiple imputation. *Ann. Statist.* **16**, 1550–66.

Titterington, D.M. & Sedransk, J. (1989). Imputation of missing values using density estimation. *Statist. Prob. Lett.* **8**, 411–8.

Wang, C.Y., Wang, S., Gutierrez, R.G. & Carroll, R.J. (1998). Local linear regression for generalized linear models with missing data. *Ann. Statist.* **26**, 1028–50.

Watson, G.S. (1964). Smooth regression analysis. *Sankyā* A **26**, 359–72.

Figure 1: (a) estimated mean squared error response surface plot for local resampling method with $n = 200$, (b) estimated mean squared error plotted against $\hat{\mu}$ for local resampling method with $n = 200$, (c) estimated mean squared error plotted against $\hat{\mu}$ for local semiparametric resampling method with $n = 200$, (d) estimated mean squared error plotted against $\hat{\mu}$ for local resampling method with $n = 100$.

Table 1: Simulation results for the first scenario. For each method: average of $\hat{\mu}$ and $S(\hat{\mu})$ (columns 1 and 2), simulated standard error of $\hat{\mu}$ (column 3), average length and estimated coverage probability of confidence intervals (columns 4 and 5) and average length and estimated coverage probability adjusted for multiple imputation (columns 6 and 7). True value is $\mu = 235.33$.

method	ave.($\hat{\mu}$)	ave.($S(\hat{\mu})$)	sse($\hat{\mu}$)	ave.CI	sim.cov.	adj. ave. CI	adj.cov.
All data	235.41	15.83	15.77	62.07	0.954	-	-
Complete cases	214.71	19.86	19.82	77.87	0.788	-	-
PSI	215.33	14.93	16.93	58.54	0.682	-	-
Rubin PMI	215.23	17.26	17.22	67.67	0.759	70.03	0.778
Efron PMI	215.12	17.08	17.38	66.94	0.739	69.07	0.755
NPSI	236.57	15.22	18.11	59.78	0.889	-	-
LSR(1)	235.86	17.58	18.13	68.96	0.925	72.39	0.925
LSR(2)	237.09	17.30	18.62	67.83	0.917	69.85	0.920
LSR(3)	236.55	17.64	18.42	69.16	0.925	72.31	0.932
LR(1)	233.53	17.38	18.71	68.13	0.919	71.97	0.924
LR(2)	234.45	17.20	18.66	67.43	0.919	69.85	0.921
LR(3)	234.20	17.52	18.87	68.69	0.916	71.97	0.920

PSI: parametric single imputation method, Rubin PMI: Rubin's parametric multiple imputation method, Efron PMI: Efron's parametric multiple imputation method, NPSI: nonparametric single imputation method, LSR: local semiparametric resampling method, LR: local resampling method. For the latter two, local weights (1) w_j, (2) $\breve{w}_j$, (3) $\ddot{w}_j$ are used.

Table 2: Simulation results for the second scenario. For each method: average of $\hat{\mu}$ and $S(\hat{\mu})$ (columns 1 and 2), simulated standard error of $\hat{\mu}$ (column 3), average length and estimated coverage probability of confidence intervals (columns 4 and 5) and average length and estimated coverage probability adjusted for multiple imputation (columns 6 and 7). True value is $\mu = 17.33$.

method	ave.($\hat{\mu}$)	ave.($S(\hat{\mu})$)	sse($\hat{\mu}$)	ave.CI	sim.cov.	adj. ave. CI	adj.cov.
LSR(1)	17.75	1.74	1.79	6.83	0.938	7.53	0.948
LSR(2)	18.24	1.72	1.94	6.75	0.906	7.30	0.918
LSR(3)	17.66	1.72	1.76	6.73	0.936	7.36	0.946
LR(1)	18.00	1.72	1.82	6.76	0.927	7.35	0.933
LR(2)	18.48	1.77	2.01	6.94	0.898	7.59	0.918
LR(3)	17.53	1.70	1.76	6.66	0.936	7.22	0.945

LSR: local semiparametric resampling method, LR: local resampling method. Local weights (1) w_j, (2) $\breve{w}_j$, (3) $\check{w}_j$ are used.