BIAS-REDUCED MULTIVARIATE IMPUTATION: USE OF THE LOCALLY-ADJUSTED PREDICTIVE MEAN MATCHING METHOD

Masato Okamoto
Statistical Research and Training Institute, Ministry of Internal Affairs and Communications
19-1 Wakamatsu-cho, Shinjyuku-ku, Tokyo-to, Japan 162-8668

Abstract

Donor imputation by the Predictive Mean Matching (PMM) method tends to yield biased estimations in multivariate cases because the distribution of potential donors selected as the donee's neighbourhood may not be centered at the donee's predictive mean. To improve the PMM imputation, I propose making adjustment of the donor's value by offsetting the difference of the (re-calculated) predicted means between the donor and the donee. The re-calculation of the predictive means is performed in the enlarged neighbourhood of the donee. This method, named the locally adjusted Predictive Mean Matching (laPMM) method, is not a complete donor imputation anymore, being in between donor and regression imputation in a sense. Empirical results based on simulation studies show a significant reduction of bias, which can be fully utilized by the fractional method for reducing MSE.

Keywords: Nearest Neighbor Imputation, Regression Imputation, Multivariate Imputation, Fractional Imputation, Predictive Mean Neighborhood

1. Introduction

The Predictive Mean Matching (PMM) method of Rubin (1986) and Little (1988) uses absolute differences between the predictive means for the recipient and potential donors estimated under a regression model, as distances between them, and chooses the nearest one or choose neighboring one at random as the donor. A generalization of PMM to multivariate imputation was proposed by Singh, Grau and Folsom (2001). Their method is termed the Predictive Mean Neighborhood (PMN) method.

However, multivariate imputation by PMM tends to yield biased estimations as shown in the following example.

Test 1

- A random sample of 10,000 observations was created from a tri-variate log-normal model with 4 covariates (x's) and random noises as below:

$$y_a = \exp(x_a + x_c + \varepsilon_1)$$
$$y_b = \exp(x_b + x_c + \varepsilon_2)$$
$$y_c = \exp(\sqrt{2}x_d + \varepsilon_3)$$

where $x_\bullet, \varepsilon_\bullet \sim N(0,1)$ i.i.d..

- 4,000 observations out of the sample were chosen to be the 'item non-response' cases at random. All outcome variables (y's) of the 'item non-response' cases were regarded to be missing.
- To impute the values for the 'item non-response' cases, an univariate log-linear imputation model was fitted for each of the 3 outcome variables. The covariates used in the models were the same as those used to create the sample. The imputation was performed using several methods including PMM, Nearest Neighbor (NN) and Probabilistic Regression (PReg) method. As for PMM, distances between the donee and the potential donors were calculated based on the Mahalanobis distance with (a) the diagonal matrix of variances of the residuals ($\log y_\bullet - \log \hat{y}_\bullet$) or (b) the variance-covariance matrix of the log-transformed predictive means ($\log \hat{y}_\bullet$) in the 3-dimensional space of $\log \hat{y}_\bullet$'s while NN were based on the Mahalanobis distance with the variance-covariance matrix of the covariates.

Table 1. Bias and MSE relative to mean (Test 1).

Method		Ave. diff.* of mean	Average* $\sqrt{\text{MSE}}$
SHD		0.0002	0.0286
PReg		0.0002	0.0194
NN		0.0160	0.0250
NN10p		0.0274	0.0332
PMM	(a)	0.0087	0.0217
	(b)	0.0094	0.0220
PMM10p	(a)	0.0179	0.0261
	(b)	0.0193	0.0270
laPMM(60)	(a)	0.0006	0.0225
	(b)	0.0008	0.0226
laPMM(60,10p)	(a)	0.0007	0.0221
	(b)	0.0008	0.0223

* $\sqrt{\sum(\text{relevant value for each } y)^2} \,/\, \text{mean of } y\text{'s total}$.

- The generation of random sample and the imputations were iterated 10,000 times. Estimates of the means of the outcome variables under PMM and NN were found to be different from the 'true' values originally generated about 0.9 % or more on average, substantially larger than those of PReg and the Simplest Hot Deck (SHD) imputation as given in Table 1.

In this paper, "SHD" refers to a hot deck method with only 1 imputation class – i.e. the sample is not stratified for the imputation. "NN10p" refers to a probabilistic NN that determines the donor at random among 10 complete observations (potential donors) nearest to the donee while "NN" corresponds to a deterministic type that chooses the most nearest one to be the donor. Similarly, "PMM10p" and "PMM" in Table 1 refer to PMM that choose the donor in a probabilistic and a deterministic manner, respectively. "laPMM(60)" and "laPMM(60,10p)" in Table 1 will be mentioned in section 3.

The above example indicates methodological improvement is required for PMM in multivariate cases. Before introduing the laPMM method, it should be noted that the imputations using (a) the diagonal matrix of variances of the residuals instead of (b) the variance-covariance matrix of the predictive means produced better estimates under PMM in the above example. The superiority of (a) stands up to the other two examples presented later. The difference between (a) and (b) may depend on data to be processed. For example, the difference may be affected by correlations among residuals; however, the type (a) seems to be superior to the type (b), on the whole, although PMN of Singh, Grau and Folsom (2001) adopted the latter type.

2. Proposed Method

Bias in estimates under PMM is considered to be brought about by the following causes.

- Potential donors selected as the recipient's neighbourhood may not be centered at the recipient's predictive mean.
- The predictive means may not be estimated with sufficient accuracy due to misspecification of imputation model such as applications of linear models to non-linear relations among variables.

The locally adjusted Predictive Mean Matching (laPMM) method is intended to reduce the bias by offsetting the difference of the predictive means between the donor and the donee, which is re-estimated by fitting simple regression of the actual values to the predictive means in the enlarged neighbourhood. The procedure for implementing laPMM can be outlined as follows.

- A given number of complete observations nearest the donee are assigned to the members of the enlarged neighbourhood using the predictive means and the PMM distance function.
- The actual values of each outcome variable are regressed on its predicted values respectively in the enlarged neighbourhood – i.e. a simple univariate linear model is fitted independently for each outcome variable using its predictive mean as a covariate. In the case that a log-linear model is used as the imputation model, values should be log-transformed for the re-calculation as follows:

$$\log y_\bullet \sim \alpha_\bullet + \beta_\bullet \log \hat{y}_\bullet$$

where $y_\bullet$: actual value, $\hat{y}_\bullet$: predicted value.

- The donor's actual value of each outcome variable is adjusted before making donation using the estimated regression parameter as follows:

$$\log y_\bullet^{\text{donor}} + \hat{\beta}_\bullet \left(\log \hat{y}_\bullet^{\text{donee}} - \log \hat{y}_\bullet^{\text{donor}} \right).$$

Since the donor's values are slightly adjusted as above, this modified PMM is not a complete donor imputation method anymore.

3. Illustrative Examples

Test 1 (continued)

Using the test setting described in section 1, laPMM was also applied to compare with PMM, NN. "laPMM(60)" in Table 1 refers to a deterministic laPMM that chooses the most nearest one to be the donor, and adjusts its values based on the relation between the actual values and the predicted values within the enlarged neighbourhood consisting of 60 complete observations nearest to the donor while "laPMM(60,10p)" corresponds to a probabilistic laPMM that determines the donor at random among 10 complete observations nearest to the donee, and adjusts its values in the same way as "laPMM(60)" does. As shown in Table 1, biases in estimates under laPMM were below 0.1%, substantially smaller than PMM and NN. This better performance of laPMM seems to be insensitive to the size of the enlarged neighbourhood if taking size 30 or over as explained later.

In terms of MSE, laPMM also produces better results than those of PMM and NN when comparing probabilistic types. In this example, the deterministic PMM is slightly better than others except PReg; however, it dose not always hold true as illustrated in the subsequent examples.

To take full advantage of laPMM's bias-reduction effect, a similar method to the Fractional Predictive Mean Matching method of Beissel-Durant and Skinner (2004) is effective. "NN10f", "PMM10f" and "laPMM(60,10f)" in Appendix 1 refer to NN, PMM and laPMM incorporated with this fractional method – i.e. 10 imputed datasets were created by randomly choosing the donor among the nearest 10 complete observations without replacement. Estimates under the imputation methods incorporated with the fractional technique were obtained by averaging 10 aggregates derived from 10 imputed datasets. As shown in Appendix 1, the fractional laPMM produced estimates with the smallest MSE, even smaller than that of PReg.

As for preservation of the univariate marginal distributions, laPMM achieve higher p-values of the

Kolomogorov-Smirnov test for each outcome variable and the total of all 3 outcome variables in comparison with NN and PMM. In Appendix 1, only the K-S test results for the total of all 3 outcome variables are listed because of space limitations.

Needless to say, good imputation must preserve relations among outcome variables and between outcome variables and covariates as well as the marginal distribution of each outcome variable. laPMM seems to produce better estimates in this respect also because biases in estimates of correlation coefficients among outcome variables and between outcome variables and covariates were smaller than those of NN and PMM (see Appendix 1). Especially the fractional laPMM achieves the smallest MSE as well as the smallest bias (except PReg in the later comparison). Some people may contend laPMM other than the fractional version are inferior to NN and PMM because of larger MSE in estimates of correlation coefficients. However, their MSE is at about the same level as that of PReg, which can be regarded as the model imputation method in this example. That is to say the increase of MSE in comparison with NN and PMM should be considered as a reasonable price paid for the reduction of bias.

Test 2

In many cases, continuous variables of actual survey data take value zero with some probability greater than zero. To check the performance of laPMM in such cases, the following example was prepared.

- A random sample of 10,000 observations was created from a tri-variate mixture model of log-normal and binomial with 5 covariates (x's) and random noises as below. Each outcome variable takes value zero with probability of about 27% (see Appendix 4).

$$z_a \sim \text{Binom}\left(p = 1/(1+\exp(-0.1(x_a + x_e + 10)))\right)$$
$$z_b \sim \text{Binom}\left(p = 1/(1+\exp(-0.1(x_b + x_e + 10)))\right)$$
$$z_c \sim \text{Binom}\left(p = 1/(1+\exp(-0.1(\sqrt{2}x_d + 10)))\right)$$
$$y_a = z_a \cdot \exp(x_a + x_c + \varepsilon_1)$$
$$y_b = z_b \cdot \exp(x_b + x_c + \varepsilon_2)$$
$$y_c = z_c \cdot \exp(\sqrt{2}x_d + \varepsilon_3)$$
where $x_\bullet, \varepsilon_\bullet \sim N(0,1)$ i.i.d..

- 4,000 observations out of the sample were chosen to be the 'item non-response' cases at random. All outcome variables (y's) of the 'item non-response' cases were regarded to be missing.
- To apply PMM and laPMM, a two-stage imputation model was fitted for each of the 3 outcome variables. Firstly, a logit model was fitted for predicting whether $y_\bullet > 0$ or not. Secondly, a log-linear model was fitted for predicting the value of $\log y_\bullet$ in the case $y_\bullet > 0$. The covariates used in the models were the same as those used to create the sample except that the following variable was added to the second stage model for the selection-bias correction.

$$\log \hat{p}_\bullet + (1-\hat{p}_\bullet)/\hat{p}_\bullet \log(1-\hat{p}_\bullet)$$

where $\hat{p}_\bullet$ is the predicted probability of $y_\bullet > 0$ obtained from the first stage model.
- As for PMM and laPMM, distances between the donee and the potential donors were calculated based on the Mahalanobis distance with the following matrix in the 6-dimensional space of $\text{logit}(\hat{p}_\bullet)$'s and $\log \hat{y}_\bullet$'s.

$$\begin{bmatrix} \mathbf{V}_{\text{logit}(\mathbf{p})} & \\ & \mathbf{S}_\mathbf{y} \end{bmatrix}$$

where $\mathbf{V}_{\text{logit}(\mathbf{p})} = \text{diag}[\text{var}(\text{logit}(\hat{p}_\bullet))]$,

$\mathbf{S}_\mathbf{y}$: (a) diagonal matrix of var. of residuals or

(b) var. - covar. matrix of $\log \hat{y}_\bullet$.

- Similar to Test 1, the generation of random sample and the imputations were iterated 10,000 times.

Due to time constraints, alternatives to the diagonal submatrix $\mathbf{V}_{\text{log it}(\mathbf{p})}$ composed of variance of $\text{logit}(\hat{p}_\bullet)$ were not pursued thoroughly; however, better alternatives have not been found so far.

- As for laPMM, the donor's actual value of each outcome variable was adjusted in the case $y_\bullet > 0$. Regression of (log-transformed) actual values of each outcome variable on the corresponding predicted values was performed in the enlarged neighbourhood excluding complete observations taking value zero.

Table 2. Bias and MSE relative to mean (Test 2).

Method		Ave. diff.* of mean	Average* $\sqrt{\text{MSE}}$
SHD		0.0005	0.0326
PReg		0.0002	0.0243
NN		0.0232	0.0330
NN10p		0.0355	0.0419
PMM	(a)	0.0142	0.0281
	(b)	0.0174	0.0296
PMM10p	(a)	0.0248	0.0339
	(b)	0.0291	0.0370
laPMM(60)	(a)	0.0008	0.0346
	(b)	0.0010	0.0336
laPMM(60,10p)	(a)	0.0006	0.0278
	(b)	0.0007	0.0275

* See the footnote to Table 1.

As shown in Table 2, results are similar to those of Test 1. Estimates of the means of outcome variables under PMM and NN were different from the 'true' values originally generated more than 1.4% on

average, substantially larger than those of SHD and PReg while, under laPMM, biases were equal or less than 0.1%. In terms of MSE, the probabilistic laPMM produces better results than those of the corresponding NN and PMM, and slightly better than that of the deterministic PMM also in this example.

The probabilistic laPMM also produces better estimates of correlation coefficients among outcome variables and between outcome variables and covariates with small biases at about the same level as those of PReg. The fractional laPMM achieves the smallest MSE retaining biases at about the same level as those of PReg (See Appendix 3). laPMM also fulfills better K-S test results for each outcome variable and the total of all 3 outcome variables. In Appendix 3, the K-S test results for the total of all 3 outcome variables are listed.

In Test 1 and Test 2, independent noises (residuals) were added to outcome variables. laPMM's superiority over NN and PMM also holds true in the cases that positively and/or negatively correlated noises are added to outcome variables. The detail has to be omitted because of space limitations. Instead, a simulation result based on actual survey data is presented next.

Test 3 using UK Family Budget Data

In the former two examples, we assume the underlying data structure to be known exactly. In reality, however, we usually have only limited knowledge about the underlying data structure, or sometimes it is too costly to apply more accurate but complex models. That is why actual survey data – UK household expenditure by major category (Blundell, Duncan and Pendakur, 1998) was used in the third example. A regression model similar to that of Test 1 or Test 2 was fitted independently for each outcome variable (expenditure for each category in this example) depending on existence of zero point mass although the residuals are correlated with each other, and the simple (log-)linear regression models are unlikely to fit completely (see Blundell, Duncan and Pendakur, 1998).

A test using budget shares of 1,519 UK households with one or two children was conducted as follows:

- As total expenditure and income of each household are rounded to the nearest 10 UK pounds sterling in the available data, a random number between -5 and 5 generated from the uniform random distribution was added to each.
- Expenditure for each category was obtained by multiplying its budget share and the total expenditure together for each household.
- 300 observations out of the sample (nearly 20% of the sample) were chosen to be the 'item non-response' cases at random. All expenditure data of the 'item non-response' cases were regarded to be missing.
- To apply PMM and laPMM, a regression model was fitted for each category independently. Since food expenditure and 'other' expenditure are always greater than zero, a model similar to Test 1 was fitted for each while a two-stage model similar to Test 2 was fitted for each of the other four categories. Instead of log-transformation, squared root of expenditure was used as the outcome variable for each category other than food in consideration of the shape of expenditure distribution. Food expenditure was not transformed. The covariates were chosen among income, age of household head and number of children as shown in Table 3. The PMM distances between the donee and the potential donors were calculated based on the Mahalanobis distance with the following matrix in the 8-dimensional space of $\text{logit}(\hat{p}_\bullet)$'s for cloth and alcohol expenditure and squared root of $\hat{y}_\bullet$'s for non-food categories and $\hat{y}_\bullet$ for food. Fuel and transport expenditure may also take value zero by a little chance; however, predicted probabilities of non-zero value were omitted from the distance space because any logit model was insignificant for both of them. The additional covariate for the selection-bias correction was not used for any category due to statistical insignificance.

$$\begin{bmatrix} \mathbf{V}_{\text{logit}(\mathbf{p})} & \\ & \mathbf{S}_\mathbf{y} \end{bmatrix}$$

where $\mathbf{V}_{\text{logit}(\mathbf{p})} = \text{diag}[\text{var}(\text{logit}(\hat{p}_\bullet))]$,

$\hat{p}_\bullet$: predicted prob. of non - zero for cloth or alcohol

$\mathbf{S}_\mathbf{y}$: (a) diagonal matrix of var. of residuals or (b) var. - covar. matrix of predicted values.

- The generation of random sample and the imputations were iterated 10,000 times. Results are summarized in Table 4 and Appendix 5.

Table 3. Covariates in imputation model (Test 3).

	income	age	Children
Food	Y (log)	Y (^2)	Y
Fuel*	N	N	N
	Y (log)	N	N
Cloth*	Y (log)	Y (^2)	N
	Y (log)	N	N
Alcohol*	Y (log)	Y	N
	N	Y (log)	Y
Transport*	N	N	N
	Y (log)	N	N
Other	Y (log)	N	N

* Covariates in both the 1st stage logit model (upper) and the 2nd stage model (lower) are listed. As for fuel and transport expenditure, only constant term was used in the 1st stage logit model.

In Test 3, PReg can not be regarded as the model method anymore due to model misspecification while

NN and PMM produced estimates with substantially small biases and MSE relative to SHD and PReg in this example, at least in the case that the type (a) amtrix is used to define the Mahalanobis distance. The probabilistic laPMM further reduced the bias slightly without an increase of MSE.

The K-S test shows the deterministic and probabilistic laPMM also preserved the univariate marginal distribution of each outcome variable and total expenditure slightly better than their counterparts respectively. In Appendix 5, the p-values for total expenditure are listed.

Table 4. Bias and MSE relative to mean (Test 3).

Method		Ave. diff.* of mean	Average* $\sqrt{\text{MSE}}$
SHD		0.00792	0.00977
PReg		118.7418	119.5496
NN		0.00045	0.00545
NN10p		0.00062	0.00548
PMM	(a)	0.00041	0.00542
	(b)	0.00478	0.02403
PMM10p	(a)	0.00052	0.00543
	(b)	0.00152	0.00983
laPMM(60)	(a)	0.00041	0.00543
	(b)	0.00351	0.01905
laPMM(60,10p)	(a)	0.00035	0.00543
	(b)	0.00055	0.00618

* See the footnote to Table 1.

As for estimates of correlation coefficients among expenditure categories, laPMM was as good as PMM while laPMM (in the case of the type (a)) was as good as NN as to those between expenditure categories and household attributes, among which covariates were chosen. Thus, Test 3 also demonstrated that laPMM was able to preserve both 'internal' and 'external' multivariate structure simultaneously. These results were retained by the fractional laPMM also.

Size of the enlarged neighbourhood

The above-mentioned comparison tests were performed with the enlarged neighbourhood of size 60 – i.e. the enlarged neighbourhood consists of 60 complete observations nearest to the donee. Results of imputations using different sizes of the enlarged neighbourhood are listed in Appendix 7 for Test 1 and Appendix 8 for Test 3. The corresponding results for Test 2 are omitted because of space limitations; however, the tendency is similar to that of Test 1.

The test results indicate bias and MSE in estimates are relatively insensitive if the size of the enlarged neighbourhood is 30 or over. In Test 1 and Test 2, the larger enlarged neighbourhood yields the smaller MSE in estimates of means and correlation coefficients. The larger size also improves bias in estimates of means while it worsens bias in estimates of correlation coefficients. In Test 3 using UK family budget data, the performances of laPMM with different sizes of the enlarged neighbourhood are at almost the same level if taking size of 30 or over; however, excessively enlarged neighbourhood deteriorates estimates slightly. Thus, the size around 60 seems an appropriate choice.

4. Final Remarks

Empirical results based on simulation studies indicate bias reduction can be expected by the use of laPMM although the effect may differ in degree depending on data to be processed and/or regression model to be applied. In the case that PMM yields nearly unbiased estimates such as Test 3, laPMM seem to be less advantageous. However, remarkable improvement may possibly be obtained in the case that PMM yields clearly biased estimates.

In some data processing tasks such as imputation of missing income by source, which is the final destination of my present research, accurate prediction of the probability of taking value zero has a great importance because several income sources reach households less frequently although incomes from the sources account for significant portions. Thus, the effects of laPMM should be reinforced with some other methods in such cases. For example, combination with the centered PMN suggested by Singh, Grau and Folsom (2004) aiming at unbiased imputation of categorical variables may be worth the consideration.

References

Blundell, R., Duncan, A., and Pendakur, K. (1998). "Semiparametric estimation and consumer demand," Journal of Applied Econometrics, 13, 435-461.

Little, R.J.A. (1988), "Missing-data adjustments in large surveys," Journal of Business and Economic Statistics, 6, 287-301.

Rubin, D.B. (1986). "Statistical matching using file concatenation with adjusted weights and multiple imputations," Journal of Business and Economic Statistics, 4, 87-94.

Singh, A.C., Grau, E.A., and Folsom, R.E., Jr. (2001). "Predictive Mean Neighborhood imputation with application to the person-pair data of the National Household Survey on Drug Abuse," 2001 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association.

Singh, A.C., Grau, E.A., and Folsom, R.E., Jr. (2004). "Imputation and unbiased estimation: Use of centered Predictive Mean Neighborhood method," 2004 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association: 4351-4358.

Appendix 1. Evaluation of imputation results (Test 1).

		KS-test[1]	Mean of y's		Correlation among y's		Correlation with x's	
		p-value	ave.[2] diff.	ave.[2] $\sqrt{\text{MSE}}$	ave.[3] diff.	ave.[3] $\sqrt{\text{MSE}}$	ave.[3] diff.	ave.[3] $\sqrt{\text{MSE}}$
SHD		0.7361	0.0002	0.0286	0.0030	0.0236	0.0681	0.0715
PReg		0.9366	0.0002	0.0194	0.0004	0.0219	0.0001	0.0244
NN		0.8308	0.0160	0.0250	0.0019	0.0197	0.0051	0.0218
NN10p		0.6582	0.0274	0.0332	0.0017	0.0193	0.0029	0.0211
PMM	(a)	0.8703	0.0087	0.0217	0.0031	0.0204	0.0061	0.0222
	(b)	0.8707	0.0094	0.0220	0.0020	0.0199	0.0061	0.0221
PMM10p	(a)	0.8312	0.0179	0.0261	0.0032	0.0200	0.0056	0.0217
	(b)	0.8273	0.0193	0.0270	0.0018	0.0195	0.0052	0.0217
laPMM(60)	(a)	0.8979	0.0006	0.0225	0.0015	0.0223	0.0024	0.0238
	(b)	0.8986	0.0008	0.0226	0.0015	0.0220	0.0021	0.0238
laPMM(60,10p)	(a)	0.9112	0.0007	0.0221	0.0007	0.0221	0.0010	0.0245
	(b)	0.9117	0.0008	0.0223	0.0009	0.0223	0.0008	0.0246
NN10f			0.0224	0.0260	0.0017	0.0182	0.0030	0.0201
PMM10f	(a)		0.0147	0.0198	0.0032	0.0185	0.0055	0.0207
	(b)		0.0157	0.0206	0.0017	0.0182	0.0052	0.0206
laPMM(60,10f)	(a)		0.0006	0.0146	0.0005	0.0174	0.0010	0.0200
	(b)		0.0007	0.0146	0.0007	0.0176	0.0008	0.0200

(a) The Mahalanobis distance based on the diagonal matrix of variances of the residuals, (b) the Mahalanobis distance based on the variance-covariance matrix of the predictive means,
1) Results of the Kolmogorov-Smirnov test applied to the total of all 3 outcome variables (y's),

2) $\sqrt{\sum(\text{relevant value for each } y)^2}\,\big/\,\text{mean of } y\text{'s total}$, 3) $\sqrt{\text{average of } (\text{relevant value for each pair})^2}$.

Appendix 2. Population means and correlation coefficients of Test 1 data.

	Mean	Correlation coefficients						
		y_a	y_b	y_c	x_a	x_b	x_c	x_d
y_a	4.4817		0.0918	0.0000	0.2314	0.0000	0.2314	0.0000
y_b	4.4817			0.0000	0.0000	0.2314	0.2314	0.0000
y_c	4.4817				0.0000	0.0000	0.0000	0.3272

Appendix 3. Evaluation of imputation results (Test 2).

		KS-test[1]	Mean of y's		Correlation among y's		Correlation with x's	
		p-value	ave.[2] diff.	ave.[2] $\sqrt{\text{MSE}}$	ave.[3] diff.	ave.[3] $\sqrt{\text{MSE}}$	ave.[3] diff.	ave.[3] $\sqrt{\text{MSE}}$
SHD		0.7354	0.0005	0.0326	0.0024	0.0239	0.0528	0.0558
PReg		0.9021	0.0002	0.0243	0.0002	0.0227	0.0003	0.0201
NN		0.7453	0.0232	0.0330	0.0013	0.0200	0.0029	0.0179
NN10p		0.5465	0.0355	0.0419	0.0013	0.0198	0.0004	0.0175
PMM	(a)	0.8141	0.0142	0.0281	0.0020	0.0206	0.0044	0.0184
	(b)	0.8041	0.0174	0.0296	0.0010	0.0201	0.0040	0.0182
PMM10p	(a)	0.7460	0.0248	0.0339	0.0019	0.0199	0.0032	0.0179
	(b)	0.6972	0.0291	0.0370	0.0009	0.0197	0.0027	0.0177
laPMM(60)	(a)	0.8649	0.0008	0.0346	0.0008	0.0233	0.0015	0.0199
	(b)	0.8680	0.0010	0.0336	0.0008	0.0237	0.0012	0.0201
laPMM(60,10p)	(a)	0.8757	0.0007	0.0278	0.0003	0.0239	0.0005	0.0205
	(b)	0.8778	0.0007	0.0275	0.0003	0.0233	0.0005	0.0207
NN10f			0.0290	0.0330	0.0012	0.0187	0.0004	0.0165
PMM10f	(a)		0.0201	0.0258	0.0020	0.0188	0.0032	0.0169
	(b)		0.0228	0.0279	0.0008	0.0186	0.0026	0.0168
laPMM(60,10f)	(a)		0.0003	0.0177	0.0003	0.0184	0.0005	0.0165
	(b)		0.0007	0.0177	0.0002	0.0183	0.0004	0.0164

See the footnotes to Appendix 1. As for the notation (a) and (b), see the text also.

Appendix 4. Population means and correlation coefficients of Test 2 data.

	Mean	Ratio of taking value zero	Correlation Coefficients							
			y_a	y_b	y_c	x_a	x_b	x_c	x_d	x_e
y_a	3.3582	0.2698		0.0665	0.0000	0.2017	0.0000	0.1968	0.0000	0.0049
y_b	3.3582	0.2698			0.0000	0.0000	0.2017	0.1968	0.0000	0.0049
y_c	3.4402	0.2698				0.0000	0.0000	0.0000	0.2852	0.0000

Appendix 5. Evaluation of imputation results (Test 3, UK family budget data).

		KS-test[1)]	Mean of y's		Correlation among y's		Correlation with x's	
		p-value	ave.[2)] diff.	ave.[2)] $\sqrt{\text{MSE}}$	ave.[3)] diff.	ave.[3)] $\sqrt{\text{MSE}}$	ave.[3) 4)] diff.	ave.[3) 4)] $\sqrt{\text{MSE}}$
SHD		0.7672	0.00792	0.00977	0.0099	0.0274	0.0386	0.0417
PReg		0.0000	118.7418	119.5496	0.2458	0.2468	0.0644	0.0670
NN		0.9797	0.00045	0.00545	0.0022	0.0205	0.0024	0.0160
NN10p		0.9801	0.00062	0.00548	0.0016	0.0213	0.0019	0.0161
PMM	(a)	0.9789	0.00041	0.00542	0.0031	0.0199	0.0028	0.0160
	(b)	0.9247	0.00478	0.02403	0.0099	0.0533	0.0049	0.0202
PMM10p	(a)	0.9810	0.00052	0.00543	0.0013	0.0207	0.0024	0.0162
	(b)	0.9322	0.00152	0.00983	0.0032	0.0269	0.0040	0.0190
laPMM(60)	(a)	0.9800	0.00041	0.00543	0.0030	0.0199	0.0025	0.0160
	(b)	0.9267	0.00351	0.01905	0.0128	0.0586	0.0034	0.0186
laPMM(60,10p)	(a)	0.9828	0.00035	0.00543	0.0015	0.0206	0.0017	0.0162
	(b)	0.9574	0.00055	0.00618	0.0034	0.0272	0.0016	0.0170
NN10f			0.00066	0.00425	0.0015	0.0168	0.0019	0.0125
PMM10f	(a)		0.00048	0.00423	0.0014	0.0165	0.0024	0.0125
	(b)		0.00150	0.00917	0.0031	0.0237	0.0041	0.0159
laPMM(60,10f)	(a)		0.00035	0.00424	0.0015	0.0164	0.0018	0.0126
	(b)		0.00055	0.00511	0.0034	0.0239	0.0016	0.0136

See the footnotes to Appendix 1. As for the notation (a) and (b), see the text also.
4) Correlation coefficients between category expenditures and household attributes not adopted as covariates in the imputation model are also included for the evaluation.

Appendix 6. Population means and correlation coefficients of Test 3 data (UK family budget data).

	Mean (£)	Ratio of taking value zero	Correlation Coefficients								
			Food	Fuel	Cloth	Alcohol	Trans-Port	Other	Income	Age	Children
Total	98.7	None	-0.479	-0.319	0.305	0.096	0.148	0.158	0.449	0.189	0.071
Food	35.2	None		0.102	-0.327	-0.122	-0.334	-0.354	-0.235	0.021	0.102
Fuel	9.0	0.002			-0.247	-0.134	-0.161	-0.133	-0.029	-0.040	-0.027
Cloth	10.6	0.063				-0.089	-0.185	-0.218	0.073	0.035	0.013
Alcohol	6.0	0.159					-0.217	-0.117	0.039	-0.143	-0.085
Transport	13.1	0.031						-0.296	0.008	0.027	-0.044
Other	24.9	None							0.153	0.026	-0.005

Appendix 7. Evaluation of laPMM with different sizes of the enlarged neighbourhood (Test 1).

		KS-test[1]	Mean of *y*'s		Correlation among *y*'s		Correlation with *x*'s	
		p-value	ave.[2] diff.	ave.[2] $\sqrt{\text{MSE}}$	ave.[3] diff.	ave.[3] $\sqrt{\text{MSE}}$	ave.[3] diff.	ave.[3] $\sqrt{\text{MSE}}$
laPMM(10)	(a)	0.8980	0.1099	2.2483	0.0056	0.0449	0.0172	0.1120
laPMM(15)	(a)	0.8999	0.0028	0.0472	0.0009	0.0237	0.0013	0.0279
laPMM(30)	(a)	0.8982	0.0012	0.0254	0.0008	0.0228	0.0013	0.0249
laPMM(60)	(a)	0.8979	0.0006	0.0225	0.0015	0.0223	0.0024	0.0238
laPMM(120)	(a)	0.8976	0.0004	0.0217	0.0018	0.0220	0.0028	0.0234
laPMM(240)	(a)	0.8975	0.0003	0.0215	0.0019	0.0220	0.0030	0.0231
laPMM(10,10p)	(a)	0.8623	0.0526	1.9790	0.0066	0.0465	0.0222	0.1167
laPMM(15,10p)	(a)	0.8983	0.0009	0.0420	0.0017	0.0248	0.0029	0.0291
laPMM(30,10p)	(a)	0.9105	0.0009	0.0296	0.0002	0.0229	0.0004	0.0259
laPMM(60,10p)	(a)	0.9112	0.0007	0.0221	0.0007	0.0221	0.0010	0.0245
laPMM(120,10p)	(a)	0.9105	0.0006	0.0214	0.0010	0.0218	0.0015	0.0239
laPMM(240,10p)	(a)	0.9108	0.0004	0.0211	0.0012	0.0216	0.0018	0.0236
laPMM(10,10f)	(a)		0.0614	2.2463	0.0069	0.0359	0.0220	0.0964
laPMM(15,10f)	(a)		0.0010	0.0414	0.0018	0.0188	0.0029	0.0234
laPMM(30,10f)	(a)		0.0008	0.0171	0.0002	0.0177	0.0003	0.0209
laPMM(60,10f)	(a)		0.0006	0.0146	0.0005	0.0174	0.0010	0.0200
laPMM(120,10f)	(a)		0.0005	0.0142	0.0009	0.0172	0.0016	0.0196
laPMM(240,10f)	(a)		0.0003	0.0141	0.0011	0.0172	0.0018	0.0194

See the footnotes to Appendix 1.

Appendix 8. Evaluation of laPMM with different sizes of the enlarged neighbourhood (Test 3, UK family budget data).

		KS-test[1]	Mean of *y*'s		Correlation among *y*'s		Correlation with *x*'s	
		p-value	ave.[2] diff.	ave.[2] $\sqrt{\text{MSE}}$	ave.[3] diff.	ave.[3] $\sqrt{\text{MSE}}$	ave.[3] [4] diff.	ave.[3] [4] $\sqrt{\text{MSE}}$
laPMM(10)	(a)	0.9806	0.00039	0.00551	0.0032	0.0202	0.0025	0.0162
laPMM(15)	(a)	0.9805	0.00039	0.00574	0.0030	0.0200	0.0025	0.0161
laPMM(30)	(a)	0.9802	0.00040	0.00542	0.0030	0.0198	0.0025	0.0160
laPMM(60)	(a)	0.9800	0.00041	0.00543	0.0030	0.0199	0.0025	0.0160
laPMM(120)	(a)	0.9800	0.00042	0.00543	0.0030	0.0199	0.0023	0.0160
laPMM(240)	(a)	0.9801	0.00043	0.00543	0.0030	0.0199	0.0023	0.0160
laPMM(10,10p)	(a)	0.9834	0.00048	0.00600	0.0023	0.0215	0.0022	0.0164
laPMM(15,10p)	(a)	0.9833	0.00026	0.00567	0.0020	0.0203	0.0018	0.0161
laPMM(30,10p)	(a)	0.9829	0.00035	0.00541	0.0017	0.0205	0.0017	0.0162
laPMM(60,10p)	(a)	0.9828	0.00035	0.00543	0.0015	0.0206	0.0017	0.0162
laPMM(120,10p)	(a)	0.9830	0.00035	0.00544	0.0014	0.0207	0.0018	0.0162
laPMM(240,10p)	(a)	0.9830	0.00035	0.00544	0.0015	0.0208	0.0020	0.0162
laPMM(10,10f)	(a)		0.00959	0.96860	0.0024	0.0177	0.0022	0.0133
laPMM(15,10f)	(a)		0.00022	0.00457	0.0020	0.0166	0.0018	0.0128
laPMM(30,10f)	(a)		0.00034	0.00424	0.0018	0.0164	0.0017	0.0126
laPMM(60,10f)	(a)		0.00035	0.00424	0.0015	0.0164	0.0018	0.0126
laPMM(120,10f)	(a)		0.00036	0.00424	0.0015	0.0164	0.0018	0.0126
laPMM(240,10f)	(a)		0.00036	0.00425	0.0015	0.0165	0.0020	0.0126

See the footnotes to Appendix 1 and 5.