Imputation by Propensity Matching

Murthy. N. Mittinty* Chacko. E†

Abstract

In large-scale surveys item nonresponse is a common phenomena. Many survey organizations use imputation to deal with missing data. Nearest neighbour imputation (NNI) has gained a lot more attention than other single imputation methods. However, in multivariate covariate situations, finding the nearest neighbour can be complicated when many variables need to be matched. In this paper we show a new application of the propensity score, which we call the nearest neighbour by propensity score (NNPS), for finding a donor for a recipient in multivariate situations. Propensity matching was originally used by Rosenbaum and Rubin (1983) in observational studies. We use propensity score for matching as it assures that the conditional distribution of the covariates given the propensity score is the same for the donors and recipients. NNPS is investigated using simulations assuming that the missing data is either missing at random (MAR) linear or MAR convex. We compare NNPS, with regression based nearest neighbour (RBNN) imputation and a new imputation method given by Murthy *et-al* (2003) called the nearest neighbour by dissimilarity (NNDM). The results indicate that matching by propensity scores seems to be a good choice for many situations, and has the advantage that it reduces the "curse of dimensionality".

Keywords. Propensity Score, Nearest Neighbour, Imputation, Covariate.

1 Introduction

In surveys, item nonresponse is a very common phenomena. Imputation is the common tool used to compensate for the missing data (Rubin, 1987; Chen and Shao, 2000, Rancourt, Sarndal, and Lee, 1994). Single or multiple data sets are created using an imputation technique. Many statistical organizations prefer single imputation to multiple imputation, in order to avoid problems caused by the multiple data sets (Marker *et-al*). In single imputation, there are various methods, one of which is based on matching and is commonly known as nearest neighbour imputation (NNI). The NNI method is used in many survey organizations like, Statistics Canada, the U.S. Bureau of Labor Statistics, and the U.S. Census Bureau (Rancourt, Sarndal, and Lee, 1994). When using nearest neighbour for imputation, the missing values are imputed under the assumption that the cases with similar covariates have similar responses.

Matching on covariates allows one to select the respondents with similar covariates to that of nonrespondents thereby reducing the nonresponse bias (Chen and Shao, 2000, Rubin and Rosenbaum, 1983, Zhao, 2004). But when the covariate space is multivariate, it is hard to find matched pairs with the same or even similar values to that of the covariates (X). Even in a simple case when all the variables are binary, there will be 2^p possible values of X where p is the dimension of X, this makes it hard to find matches that are homogenous in X (Rosenbaum, 2002). An alternative in such situations proposed in this paper is to use the propensity matching previously used in a different context by Rosenbaum and Rubin (1983)(RR in the rest of the paper). As defined by RR, a balancing score $b(X)$ is a function of the observed covariates such that the conditional distribution of X given $b(X)$ is the same for missing (nonrespondents, denoted by $m = 1$) and non missing cases (respondents, $m = 0$) and this is denoted by $X \perp\!\!\!\perp m|b(X)$. It is this property that allows the distribution of covariates in respondents and nonrespondents to be similar when matched using covariates or propensity scores, thereby reducing the nonresponse bias. According to RR, matching by covariates provides the finest balancing score, matching on propensity score provides the coarsest balancing score.

Use of propensity scores in missing data was first introduced by Little (1986) for forming strata prior to imputation. Later its use was shown by Lavori (1995) for

*Department of Mathematics and Statistics, University of Canterbury, New Zealand, email: nmi13@student.canterbury.ac.nz

†Department of Mathematics and Statistics, University of Canterbury, New Zealand

multiple imputation by approximate bayesian bootstrap. Both these studies use propensity score for stratifying the data prior to imputation. But in this paper we propose using propensity scores as a method for reducing the dimension of matching variables for imputation. To investigate its efficiency, we compare this method with a NNI method that uses the actual covariates, and a NNI method that uses regression for reducing the dimension of the multivariate space. All the three imputations are carried out on the data which has missingness in one variable.

The rest of the paper is organized as follows. Section 2 describes the nearest neighbour imputation method in general. In section 3 we introduce propensity matching method. Section 4 presents the NNI methods used for comparisons. In section 5 we present the assumptions and details of simulations. The results of the study are given in section 6. We conclude in section 7 that when there are only a few covariates and cases NNI by dissimilarity is the preferred method, but for a large number of covariates, matching by NNI by dissimilarity may be too slow and hence matching by propensity score or regression based nearest neighbour is preferred.

2 Nearest neighbour imputation

To begin with, let us take a simple case and illustrate matching on covariates. Consider a bivariate sample (X_1,Y_1),(X_2,Y_2).....,(X_n,Y_n). Let the variable X be covariate data that is completely observed on all n cases and let Y be observed only for n-r cases. If for any X_j, corresponding to the missing Y_j, we would have an exact match if we can find some X_i corresponding to known Y_i such that $X_i = X_j$. If, as in the general case an exact matching of the covariates corresponding to the missing observations and observations with complete response is infeasible, we weould use the method of nearest neighbour (NN). In Nearest neighbour we match X_j with in the neighbourhood of X. For imputing the missing Y_j, $j = r+1, ...n$, the NN method finds the nearest neighbour using some distance measure. If the distance d_{ij} on the observed X-variables is defined as

$$d_{ij} = |X_i - X_j|, \tag{1}$$

the nearest neighbour obtained for the missing case j is the case k where $d_{kj} = \min_{1\leq i\leq r}(d_{ij})$. In the case where X is multivariate and continuous, one might think of using a distance measure such as Mahalanobis distance. The Mahalanobis distance matching can present problems; for example when a covariate X_i is binary, the Mahalanobis metric may try hard to match this X_i exactly thus reducing the quality of match of the other covariates (Rosenbaum and Rubin, 1983). Another method that is used when the covariate space is multivariate is the NNI method that uses regression (RBNN) (Little and Rubin, 2002). However RBNN methods are not appropriate when the model assumptions are not satisfied (Allison, 2001). In situations where X has different types of variables, Murthy *et-al* (2003) has described a new matching procedure called the nearest neighbour by dissimilarity (NNDM). This takes care of different types of variables, provides efficient matching and preserves the distributions. However when there are many variables, matching on all variables is computationally intensive. Hence, we now propose the use of propensity score for matching. Rosenbaum and Rubin (1983), in the context of observational studies, have shown that propensity matching can effectively balance binary covariates for which matching is not possible on an individual basis. Here we apply this concept to imputation. Since the propensity score effectively represents all variables, we suppose that the use of it for dimension reduction in imputation might be more effective than transforming the multivariate space to univariate space by regression. This has been verified by simulations in section 5.

3 Matching by Propensity score

3.1 Propensity score

The propensity score is defined as follows by Rosenbaum (2002): Let m be the missing indicator defined on Y (the response variable) i.e if Y is observed then $m = 0$ else $m = 1$. The propensity score $\pi(X)$ is defined as $\pi(X)= Pr(m = 1|X)$.

The theory of propensity score given by RR for observational studies and later discussed in context of survey nonresponse by Little (1986) shows that if the missing data are missing at random (MAR) given X then they are missing at random given $\pi(X)$, that is if $m \perp\!\!\!\perp Y|X$ then $m \perp\!\!\!\perp Y|\pi(X)$, where $m \perp\!\!\!\perp Y|X$ means that m is conditionally independent of Y given X (for a proof, see p.48, Little and Rubin, (2002)). In other words, conditioning on the propensity score would remove the correlation between X and m, and hence replacing X with $\pi(X)$ does not lead to any loss of information because $X \perp\!\!\!\perp m|\pi(X)$ (Imbens (2004); Cook (1998)).

The response propensities ($\pi(X)$) are estimated by using logistic regression of m on X. The procedure for computing $\pi(X)$ is described next.

3.2 Computation of propensity score

As in RR we use the following form to estimate the propensity score

$$\pi(X^*) = \frac{e^{\beta X^*}}{1 + e^{\beta X^*}} \tag{2}$$

where $X^* = (1, X_1, X_2, ..., X_p)$ and $\beta = (\beta_0, \beta_1, ...\beta_p)$. The glm function in R is used to estimate $\pi(X)$, by regressing m against the covariates X, with the family binomial and link function logit. We do not use the alternate method of discriminant analysis procedure described in Rosenbaum(2002) for estimating the propensity scores because we want a method that deals with a covariate set that has mixed types of variables like binary, nominal, continuous and ordinal variables.

3.3 Nearest neighbour by propensity scores (NNPS)

Having described how we compute the propensity score, we now state some basic ideas behind propensity matching given by Rosenbaum (2002) and then define the distance measure to find the nearest neighbour.

Proposition given in Rosenbaum (2002). *If* $\pi(x) = \pi$, *then*

$$pr[X = x|\pi(X) = \pi, m = 1] =$$

$$pr[X = x|\pi(X) = \pi, m = 0] =$$

$$pr[X = x|\pi(X) = \pi]$$

The above property, known as the balancing property, states that when the propensity score π is the same for the missing and non missing cases then the distribution of covariates is the same for missing and non missing cases. The proposition considered is a particular case (where the number of strata is only one) of the original proposition given in Rosenbaum (2002), hence we avoid the proof.

For NNPS we use equation (1) to determine the nearest neighbour, but with $|X_i - X_j|$ replaced by $|\hat{\pi}(X_i) - \hat{\pi}(X_j)|$, where $\hat{\pi}(X_i)$ is the estimate of $\pi(X_i)$. We use this matching method since the distribution of covariates is the same for missing and non missing cases within the neighbourhood of $\hat{\pi}$. As $\hat{\pi}(X)$ is continuous we expect no ties among the donors obtained for imputing the missing values.
The performance of imputation by propensity matching is next compared to two other nearest neighbour imputation methods in terms of computational time and bias.

4 Methods used for comparison

As explained in section 2, there are several methods for finding the distance in the multivariate case. In this paper we use two procedures; nearest neighbour based on regression (RBNN) and nearest neighbour by dissimilarity (NNDM). These two procedures are described in the following two subsections.

4.1 Regression Based Nearest Neighbour (RBNN)

The RBNN procedure was initially given by Laaksonen (2000). RBNN is similar to the predictive mean matching method given in Rubin (1987). Under this procedure imputation is carried out in the following manner where we use the subscripts obs (mis) to refer the observed (missing) cases;

1. Using the observed cases construct a regression model
$$y_{obs} = \alpha + \beta X_{obs}$$
2. Use the estimates $\hat{\alpha}, \hat{\beta}$ of α, β to predict y (by $\hat{y}$) for all available X
$$\hat{y} = (\hat{y}_{obs}, \hat{y}_{mis}) = \hat{\alpha} + \hat{\beta}X + \epsilon$$
where ϵ is as defined below.
3. Find nearest neighbour to $\hat{y}_{mis}$ from $\hat{y}_{obs}$ for each missing case.
4. Use the y_{obs} corresponding to the nearest $\hat{y}_{obs}$ as the imputed value for y_{mis}

The error (ϵ) is assumed normally distributed (0,σ^2). It is not uncommon to assume $\sigma^2 = 0$, but this does not add variability due to imputation. If the data sets are small and the error term is added it make a difference to the final data (Allison, 2001). There are several ways to add this error term to the model (Kalton and Kaspryzk, 1987), but we use the method of Laaksonen (2000) where σ^2 is the residual mean square error estimated from the regression model.

4.2 Nearest neighbour by dissimilarity matrix (NNDM)

NNDM is the other method used for comparison as it provides good matching when the variables are of mixed type and it preserves the distributions as shown in Murthy *et-al* (2003).
The dissimilarity matrix between a complete case c and a missing case m is defined as

$$D(c,m) = \frac{\sum_{j=1}^{p-1} \delta_{cm}^{j} d_{cm}^{j}}{\sum_{j=1}^{p-1} \delta_{cm}^{j}} \tag{3}$$

where the distance d_{cm} for the j^{th} covariate (d_{cm}^{j}) is given by

$$d_{cm}^{j} = \begin{cases} \left.\begin{array}{ll} 1 & \text{if } x_{cj} \neq x_{mj} \\ 0 & \text{otherwise} \end{array}\right\} \text{for binary nominal variables.} \\ \frac{|x_{cj} - x_{mj}|}{r_j} \quad \text{for interval and ordinal variables} \end{cases}$$

δ_{cm}^{j} is an indicator variable which is 1 except when the j^{th} covariate is asymmetric and $x_{cj} = x_{mj} = 0$, and r_j is the range of the j^{th} covariate.
We use r_j rather than standard deviation to normalize as this ensures that for interval and ordinal variables $d_{cm}^{j} \in$ [0,1] as for the other binary and nominal variables. For further details on asymmetric variables refer to Kaufman and Rousseeuw (1990) or Murthy*et-al* (2003).
Of possible donors the case c with $min[D(c,m)]$ is considered the nearest neighbour and used as the donor. If there are several donors with the same $min[D(c,m)]$, a donor may be randomly selected from among them.

5 Simulation

For comparing the closeness of the imputed values by proposed matching methods, we use Monte Carlo simulations. In these simulations we assume that; a) Missingness is in one variable, b) the sample is drawn by simple random sampling, c) there is only one imputation class and, d) the missing data is MAR linear or MAR convex. The Monte Carlo simulations were performed on three different data sets.

5.1 Data generation

We used two real life data sets and a simulated data set for these comparisons. The two real life data sets used are; Tooth Growth data (TGD) (Mc.Neil,1977) and Low birth Weight data (LBW) (Hosmer and Lemeshaw, 2002). As these data sets had sample sizes of 60 and 189 respectively, to investigate the performance of these methods on large data sets, we simulated an artificial population with the covariate (X) being multivariate. In this artificial data set (AD), we have three randomly generated covariates X', Z, W, of size $N = 10,000$ and a response variable Y. In the artificial data, the variable Z is a binary variable and W is a categorical variable with three categories. The covariate X' and the response variable Y have a joint distribution $\mathrm{N}(\mu, \Sigma)$, where $\mu = (10, 12)$ and $\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 2.5 \end{bmatrix}$. All the three data sets initially have no missing values and missingness is induced into the data using MAR linear and MAR convex mechanisms. These mechanisms are detailed in the following subsection.

5.2 Missing data mechanism

The different types of missingness used in comparisons are MAR linear, and MAR convex as defined in Collins *et-al* (2001). The MAR mechanism is

1. MAR linear if the probability of missingness is linearly related to one of the covariates.

2. MAR convex if missingness is more on the extremes of the covariate and smaller in the middle

For MAR linear, missingness is created by dividing a continuous covariate into four groups based on quartiles and then set different probabilities in a linear manner such that we achieve a desired percent of missingness. For MAR convex missingness is created by first dividing a covariate into four groups based on the quartiles. The probabilities of nonresponse are set high in the first and last quartile and low in middle two quartiles to achieve a desired amount of missingness. In these simulations a covariate which is of interval type is used to form the groups.

5.3 Creating missing data

In all the three data sets missingness is induced using MAR linear and MAR convex mechanism as described in sec 5.2.
Missing data in artificially generated data (AD): For each simulation a sample of size 1,000 is drawn for the population of size 10,000 using simple random sample with replacement. We used sample with replacement in order to be consistent with the sample selection process of

the tooth growth and low birth weight data simulations where, because of the small data length, we took sample with replacement. We call this sample data set as "ADS". In order to achieve 25 and 40 percent missing rates in ADS we set the probabilities to (0.1,0.2,0.3,0.4) and twice these values under MAR linear. Similarly we set the probabilities to (0.4,0.1,0.1,0.4) and twice these values to achieve 25 and 40 percent missing data under MAR convex mechanism. X' variable of AD was used to form the quartiles.

Missing data in tooth growth data (TGD): The data set TG has three variables and is of size 60. the three variables are "length", "dose", and "supplement" (supp). Variables length and dose are of interval type and supp is binary. Length and dose are correlated the correlation being 0.80. Instead of using an interval variable to form quartiles we used a binary variable by doing so we created a special case where MAR linear and MAR convex are the same. Now in this data the probabilities were set to (0.2 and 0.4) to have around 28 percent missing data in all the simulations.

Missing data in low birth weight (LBW): This data set has an asymmetric binary variable (Smoking). The variable Age was used for generating the missingness. The probabilities for MAR linear were (0.1,0.2,0.3,0.4) for 18% missingness and (0.2,0.4,0.6,0.8) for 35 % missingness. For MAR convex the probabilities were (0.4,0.1,0.1,0.4) and (0.8,0.2,0.2,0.8) to achieve 18 and 35 percent missingness. 1,000 simulations were carried out. We used 9 out of 11 variables in our simulations omitting the variables "Case ID" and "Low". The variable "Low" in the LBW data is a categorical variable constructed from birth weight information. As we are using birth weight as a continuous variable, the variable Low is redundant. This data set has 189 observations.

In order to compare the performance of the procedures we used the mean square errors (MSE). The mean square error is computed as the squared difference between the original value before missingness is inserted and the imputed value; that is

$$\Delta = \frac{\sum_{i=1}^{n_{imp}} (y_i^{imp} - y_i^{actual})^2}{n_{imp}}$$

where y^{imp} is the imputed value and y^{actual} is the actual value of y before inducing the missingness and n_{imp} is the number of imputed values.

6 Results

Simulation Set 1:
For the first set of simulations we used ADS data. For the RBNN method the regression model had an R^2 value in between 0.43-0.79. The error term ϵ is generated with normal distribution $(0, \sigma^2)$, where σ^2 is the residual mean square error obtained from the regression as defined in sec.4.1.

The imputation results presented in Tables-1, show that NNDM imputes the missing values close to the true values in all cases. Comparison of NNPS and RBNN shows that RBNN has lower MSE under MAR convex. For MAR linear the NNPS is the preferred choice.

Simulation Set 2:
In this simulation set we used the tooth growth data. For this data the details of the regression model used in RBNN are; R^2 lies in the interval 0.65-0.72.

Imputation results are presented in Table-2. For this data the NNPS and NNDM imputes the missing values close to the true values.

Simulation set 3:
For the third set of simulations we used LBW data. Results of the regression model used in RBNN show that R^2 lies in between 0.2 and 0.45. From Table 3 it is observed once again that NNPS may be a better option than RBNN. The MSE of NNPS lies in between that of RBNN and NNDM. The MSE presented in Tables 1,2, and 3 is the mean of the MSE's obtained from 1000 simulations. The standard error reported in the tables is obtained using the bootstrap function in the R package.

Computational Time:
Table-4 presents the computational times for all the methods. The comparisons of the computational times show that there is a notable reduction in time when imputation is performed using NNPS or RBNN. For the TGD simulation, the computational time for a single simulation run by NNDM is 0.26 seconds and by RBNN and NNPS 0.03 seconds. When the number of cases were increased to 1,000 as in the ADS data, the computational time for NNDM is 29.55 seconds for MAR linear and 25% missingness, when the percent missing is around 40 the computation time increased to 73.6 seconds. For RBNN and NNPS the computational times are 0.34 and 0.50 with 25% MAR linear missingness and 0.56 and 0.88 with 40% missingness. MAR convex also gave similar results.

Table-1: Comparison of MSE under MAR linear and MAR convex with different % of missingness in simulated data.

% Missing	MAR	Matching by	MSE	SE
25	Convex	NNDM	10.91	0.012
		NNPS	11.46	0.013
		RBNN	11.00	0.013
	Linear	NNDM	10.85	0.008
		NNPS	10.96	0.009
		RBNN	11.08	0.010
40	Convex	NNDM	10.96	0.011
		NNPS	11.57	0.010
		RBNN	10.99	0.010
	Linear	NNDM	10.86	0.008
		NNPS	10.87	0.009
		RBNN	11.01	0.010

Table-2: Comparison of MSE for the Tooth Growth data with 28 percent missing data.

Matching by	Missing Mechanism	
	MAR	
	MSE	SE
NNDM	6.39	0.002
NNPS	7.06	0.004
RBNN	15.27	0.006

Table-3: Comparison of MSE under MAR linear and MAR convex with different % of missingness in Low birth weight (LBW) data.

% Missing	MAR	Matching by	MSE	SE
18	Convex	NNDM	0.334	0.002
		NNPS	0.38	0.005
		RBNN	0.38	0.005
	Linear	NNDM	0.22	0.003
		NNPS	0.24	0.005
		RBNN	0.25	0.004
35	Convex	NNDM	0.33	0.003
		NNPS	0.36	0.006
		RBNN	0.40	0.006
	Linear	NNDM	0.34	0.003
		NNPS	0.40	0.004
		RBNN	0.46	0.009

Table-4 Computational time in seconds for the three methods under different percent of missingness.

Data set	% missing	Matching by		
		NNDM	NNPS	RBNN
AD	25	29.5	0.34	0.50
	40	73.6	0.56	0.88
LBW (linear)	18	1.53	0.03	0.02
	35	1.84	0.05	0.03
LBW (convex)	18	0.84	0.03	0.02
	35	1.42	0.03	0.02
TGD	28	0.26	0.03	0.02

7 Conclusions

From the results obtained, we observed that, when there are few covariates and cases the use of all covariates to find nearest neighbour is recommended; when the covariates are of different types NNDM should be used. When the number of covariates and/or the number of cases increase, NNDM is still more accurate but may be too slow. In these cases the NNPS method is better if the missingness in data is MAR linear and the RBNN method is better if the missingness in data is MAR convex. However since this is based on a limited number of situations, further study will be needed to confirm these findings. In particular we intend to apply these methods to the National Family Health Survey data and investigate their effectiveness.

References

Allison, P.D. (2001). *Missing Data*. Series: quantitative applications in the social sciences. Sage Publications: Thousand Oaks.

Chen, J., and Shao, J. (2000). Nearest neighbour imputation for survey data. *Journal of official statistics*: Vol 16, No. 2, pp 113-131.

Collins, L.M., Schafer, J.L., and Kam, C.M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*: Vol 6, No.4, pp 330-351.

Cook, R.D. (1998). *Regression Graphics: Ideas for studying regression through graphics*, John Wiley: New York.

Hosmer, D.W., and Lemeshaw, S. (2000). *Applied logistic regression*. 2^{nd} edition. John Wiley: New York.

Imbens, G.W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review, *The Review of Economics and Statistics*: Vol 86, No.1, pp 4-29.

Kalton, G., and Kaspryzk, D. (1981). Imputing for survey response. *Proceedings of the section on survey research methods*. American statistical association. pp 22-33.

Kaufman, L., and Rousseeuw, P. J. (1990). *Finding Groups in Data. An introduction to cluster analysis*. John Wiley: New York.

Laaksonen, S. (2000). Regression based nearest neighbour hot decking, *Computational statistics*: Vol 15, No.1, pp 65-71

Lavori, P.W., Dawson, R., and Shera, D. (1995). A Multiple imputation strategy for clinical trials with truncation patient data, *Statistics in Medicine*: Vol 14, pp 1913-1925.

Little, R.J.A.(1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*: Vol 54, pp 139-157.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing data*. 2^{nd} edition. John Wiley: New York.

Marker, D.A., Judkins, D.R., and Winglee, M. (2002). Large scale imputation for complex surveys. *Survey Nonresponse*. Ed. Groves, R.M., Dillman, D.A., Eltinge, J.l., and Little, R.J.A. John Wiley: New York. item Mc.Neil, D.R. (1977). *Interactive Data Analysis* Wiley: New York.

Murthy, M.N., Chacko, E., Penny, R., and Monir Hossain, Md. (2003). Multivariate nearest neighbour imputation. *Journal of Statistics in Transition*: Vol 6 , No.1 , pp 55-66

Rancourt, E., Sarndal, C., and Hyunshik Lee. (1994). Estimation of the variance in the presence of nearest neighbour imputation.*Proceedings of the section on survey research methods*. American Statistical Association. pp 888-893.

Rosenbaum, P.R. (2002). *Observational studies*. 2^{nd} edition. Springer-Verlag: New York.

Rosenbaum, P.R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*: 70, No. 1, pp 41-55.[Denoted as RR in the paper]

Zhao, Zhong. (2004). Using matching to estimate treatment effects data requirements, matching metrics, and monte carlo evidence, *The Review of Economics and Statistics*: Vol 86, No.1, pp 91-107.

Acknowledgements

The authors wish to thank Richard Penny, and Jana Asher for their comments and help in improving this paper. Earlier version of this paper presented at JSM 2004,Toronto.