

A tree based procedure for multivariate imputation

Una Procedura ad Albero per l'Imputazione Multivariata

Riccardo Borgoni

Southampton Statistical Sciences Research Institute,
University of Southampton,
borg1@socsci.soton.ac.uk

Ann Berrington

Division of Social Statistics,
University of Southampton,
amb6@socsci.soton.ac.uk

Riassunto: In questo lavoro proponiamo una procedura di imputazione multivariata basata sull'applicazione iterata di sequenze di modelli ad albero. Un'applicazione al rischio di essere fumatori esemplifica la procedura proposta e mostra come questa possa essere integrata con metodi di tipo bootstrap e di imputazione multipla per stimare la variabilità delle stime tenendo conto del processo di non risposta.

Keywords: Classification Tree, Multivariate Imputation, Bootstrap, Multiple Imputation

1. Introduction

In this paper we propose a tree based approach for multivariate imputation. We demonstrate through a case study how computer intensive methods can be used to address the extra variability in regression coefficient estimates due to the procedure¹.

A tree model (Breiman et al 1984) provides the conditional distribution of an outcome variable Y given a vector of predictors X . Those models are fully described by the pair (T, θ) where T is a tree with R terminal nodes and $\theta = \{\theta_1, \dots, \theta_R\}$ is a vector of parameters such that θ_r is associated to the leaf r for $r=1, \dots, R$. If X lies in the region corresponding to the leaf r , then $Y|X$ is distributed according to a probability law $p(Y|\theta_r)$. A binary tree partitions the predictor space into R subsets identified by the terminal nodes of the tree. A splitting rule is applied to each internal node using the predictors to allocate every unit to its left or right child node. We consider the case where both the predictor and all of the regressors are categorical, which is a common situation in applications based on survey data. In the categorical predictor case, the splitting rule classifies a subject i into, say, the left child node if $\{X_i \in C\}$ C being a subset of categories. For a categorical response variable, each terminal node identifies a multinomial law of parameter θ_r where θ_r is the vector of multinomial probabilities $\theta_{rj} = \Pr\{y=j|r\}$, $j \in \mathcal{J}$, $\mathcal{J}$ being the set

¹ We thank Ray Chambers for his comments. Any errors remain the fault of the authors. We acknowledge the financial support of the Department of Health for England grant number 0370020. The views expressed here are those of the authors and not necessarily the Department of Health. Data from the 1970 British Cohort Study were made available via the UK Data Archive.

of categories of Y , which may differ for different terminal node r . The splitting procedure moves towards more homogenous nodes and the tree is usually tailored in such a way that some classification cost criterion is minimized.

The use of classification trees for multivariate imputation is a relatively recent development. Previous studies in this context were provided amongst others by Barcena and Tussel (2002) and Mesa et al. (2000). Multivariate imputation using sequences of regressions are considered by Raghunathan et al. (2001). In this paper we present an imputation procedure consisting of fitting a sequence of classification trees. For each grown tree the imputation is obtained by drawing values from the conditional distribution identified by the leaves. The procedure is then iterated until convergence. In the next session we describe the algorithm used, whilst section 3 presents a case study which exemplifies the imputation procedure and methods for variance estimation.

2. An iterative tree based procedure

Given a sample of size n let $\mathbf{Y} = \{Y_1, \dots, Y_s\}$ be a set of s variables which are affected by non-response. For each variable Y_k , $k=1, \dots, s$, let Y_k^O be the subvector of the observed individuals and Y_k^M the residual part of missing observations. Let $\mathbf{X} = \{X_1, \dots, X_q\}$ be a vector of q fully measured variables. Let $O_k = \{Y_k^O, X^{Y_k^O}\}$ be the subsample of those for whom Y_k is observed and $M_k = \{Y_k^M, X^{Y_k^M}\}$ the subsample of those for whom Y_k is missing. For each Y_k that needs to be imputed, we grow a tree, T_k , on the data set O_k . We then apply the obtained tree to the data set M_k and make a prediction for the missing values of Y_k . Usually a prediction is obtained via a tree by dropping a new unit in the tree and labeling it according to the modal category in the leaf that the unit reaches. This implies that all the units which belong to the same terminal node are assigned the same predicted value. A different approach can be used instead. For all of the units in M_k that reach the same leaf, a random draw is conducted from a Bernoulli or a multinomial distribution, depending on whether the variable of interest is dichotomous or polytomous, estimated in each terminal node. In each step of the procedure a data set which includes variables imputed prior to the current step is passed to the tree. Once all the variables $\mathbf{Y}$ are imputed, the procedure is iterated using the same sequence of sets $\{(O_k, M_k): 1, \dots, s\}$, until a given stopping rule is met. The stopping rule can be stated in terms of a maximum number of iterations or according to stability criteria. The latter is used in the present paper. In particular the procedure stops when the distance of the probability distribution of each variable does not differ more than a given threshold ε from the same marginal distribution found in the previous step. In particular in this paper a sup-norm is used: $\max_{k=1, \dots, s} (d_k) < \varepsilon$, $d_k = \max_{j \in \mathfrak{I}_k} \{p_j^{kt} / p_j^{k(t-1)} - 1\}$ where p_j^{kt} is the

estimated probability of cell j in iteration t , $j \in \mathfrak{I}_k$, $\mathfrak{I}_k$ being the set of categories of Y_k and ε is fixed a priori. Although having fully observed covariates is not strictly necessary when the first sequence of trees is estimated and missing values in the predictor set of each tree may be managed according to a number of different strategies, our implementation begins by imputing the modal class to missing items of each variable before performing the first step.

3. Teenage motherhood and smoking in adulthood: a case study

In this section the proposed procedure is applied to an observational study investigating the relationship between teenage motherhood and regular smoking in adulthood. Data come from the 1970 Birth Cohort Study which has followed up all children born in Britain between 5-11 April 1970. Characteristics of the cohort members (CM) are collected at a number of different time points using a variety of survey instruments. There is, therefore, a complex pattern of item non-response among our sample of 4650 women. In this example we use covariates from the birth, age 10 and age 30 surveys. Apart from ethnicity, all of the covariates suffer some degree of item non-response, ranging from 0.05% to 20%.

In order to assess the impact of the imputation on the logistic regression estimates for the predictors of smoking at age 30 we compare the parameter estimates based on the complete cases ($n=3219$) with the imputed sample. In order to address the additional variability due to the imputation, we combine our algorithm with two procedures for variance estimation, a naïve form of multiple imputation (MI) and the bootstrap. One of the advantages of the proposed procedure is that we can use a number of variables for imputing items, which may or may not be fully observed, in addition to those actually used in the substantive model². For each of the considered variables affected by partial non-response a tree is grown whose size is determined via a 20 to 40-folder cross validation (the number of sub-sample actually used was determined through preliminary analyses). The optimal tree in terms of cost-complexity is then used for the imputation. The sequential procedure is iterated until the stopping rule described above is met for $\varepsilon=0.025$ ³.

For the MI we obtained 50 different filled versions of the dataset using a random generation starting from 50 different seeds. The bootstrap is based on 500 replicates of the original sample. MI estimates and their standard deviation (Little and Rubin (2002) pg. 85-89) are reported in column 4 and 5 of Table 1. Columns 6 and 7 of Table 1 show the bootstrap estimates and their standard errors. Both the procedures are computationally heavy also as a consequence of the cross validation step performed in each tree of the sequential imputation. Table 1 shows the final parsimonious logistic regression model for the predictors of smoking in adulthood⁴. The CM's age at motherhood is strongly related to the probability of being a regular smoker in adulthood, with younger mothers more prone to smoke. Apart from ethnicity, the imputation does not significantly affect the estimates of the parameters. It is interesting to note that ethnicity was the only variable not to suffer from item non-response. We suggest that there is an association between ethnicity and the level of item non-response in the remaining covariates. The standard errors obtained by the bootstrap and MI tend to be narrower than those obtained from the complete case analysis because of the

² For the imputation procedure the CM's father social class and the parents' educational level are used in addition to the variables used for the substantive model.

³ It may be noticed that the procedure managed to converge in less than 20 iterations in each of the bootstrap and MI replicates of the dataset performed in this section.

⁴ The reference profile is "not mother" (age at motherhood), "not white", "20-24" (CM's mother's age at motherhood), "< 1st quartile" (reading score), "normal conduct at age 10, benefit at age 10, satisfied with area where living, currently not in social housing and not in an unemployed family.

larger sample size. Generally speaking, MI tends to provide narrower standard errors than the bootstrap, as would be expected.

This analysis has shown how our iterative tree procedure can be usefully applied to a substantive problem. In comparison with other tree-based methodologies mentioned above, our approach is much simpler. To perform, all that is necessary is a tool to estimate a univariate tree and to loop over the variables one wants to use. An in depth comparison of our approach to other tree-based or regression-based iterative algorithms is beyond the scope of this paper but is a topic for future research.

Table 1: *Logistic regression, smoking at age 30: complete case, bootstrap and MI*

Variables	Complete Case		MI		Bootstrap	
	Coef.	Std.Er.	Coef.	Std.Er.	Coef.	Sd.Er.
Intercept	-0.96	0.16	-1.11	0.14	-1.10	0.14
Age at motherhood: <18	0.56	0.15	0.62	0.12	0.61	0.13
Age at motherhood: 18-19	0.19	0.13	0.25	0.10	0.25	0.10
Age at motherhood: 20-24	-0.11	0.12	-0.11	0.10	-0.10	0.10
Age at motherhood: >24	-0.65	0.19	-0.72	0.16	-0.74	0.16
Ethnicity: White	-2.11	0.40	-1.44	0.25	-1.48	0.26
CM's Mother's age at moth. <20	0.29	0.10	0.33	0.09	0.32	0.09
CM's Mother's age at moth. >25	-0.27	0.11	-0.27	0.09	-0.27	0.10
Reading Ability Score: 1 st -2 nd quart.	0.17	0.12	0.16	0.11	0.16	0.10
Reading Ability Score: 2 nd -3 rd quart.	-0.13	0.12	-0.09	0.11	-0.08	0.11
Reading Ability Score: > 3 rd quart.	-0.35	0.13	-0.31	0.11	-0.31	0.10
Conduct at Age 10: antisocial	0.53	0.15	0.58	0.13	0.57	0.13
Conduct at Age 10: mixed	0.29	0.28	0.37	0.23	0.35	0.26
Conduct at Age 10: neurotic	-0.33	0.18	-0.37	0.15	-0.38	0.14
Benefit in childhood: No	-0.26	0.12	-0.18	0.10	-0.19	0.10
Unsatisfied with area where living	0.38	0.12	0.40	0.10	0.40	0.10
Currently in social housing	0.67	0.13	0.68	0.10	0.69	0.10
Currently in no work family	0.43	0.14	0.45	0.11	0.45	0.12

References

- Barcena Ruiz, M. J., Tussel P. F., (2002) - Multivariate Data Imputation using Trees, Universidad del Pais Vasco, *Departamentos de Econometria y Estadistica*, Hacienda Publica y el Instituto de Economia Publica.
- Breiman, L., Friedman, J. Olshen, R., Stone, C. (1984) *Classification and Regression Trees*, Wadsworth.
- Little, R. J. A., Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, second edition, John Wiley & Sons, New York.
- Mesa, D.M., Tsai, P., Chambers R.L. (2000). Using Tree-Based Models for Missing Data Imputation: An Evaluation Using UK Census Data. *AUTIMP*.
- Raghunathan E. T., Lepkowski, J. M., Van Hoewyk, J., Solenberger, P. (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models, *Survey Methodology*, 27, 85-95.