Paper SD113

A SAS® Algorithm for Imputing Discrete Missing Outcomes Based on Minimum Distance

Macaulay Okwuokenye, Biogen Inc, Cambridge, MA, USA and Karl E. Peace, Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA, USA

ABSTRACT

Missing outcome data are encountered in many clinical trials and public health studies and present challenges in imputation. We present a simple and easy to use SAS-based imputation method for missing discrete outcome data. The method is based on minimum distance between baseline covariates of those with missing data and those without missing data. The imputation algorithm, a method that may be viewed as a variant of the hot dec imputation method, imputes missing values that are "close to" the observed values, implying that had there been data on those missing, it would have been similar to those non-missing. An illustrative example is presented.

Keywords: Mahalanobis Distance, missing discrete data, imputation, hot dec, count data, imputation of MRI data

INTRODUCTION

Missing data in clinical trials and public health studies are almost unavoidable when studies are conducted in human subjects due to possibility of participants not being available for assessment of treatment outcome. The possible reasons for missing data are numerous: it could be related to a participant's disease status e.g., improvement or worsening of symptoms or death; it could also be completely unrelated to disease status or treatment, e.g., a participant could skip clinic visit due to inclement weather, say.

Missing data impact statistical analysis and interpretation of results. Besides reducing statistical power, missing data could result in biased estimates [1]. In the case of a randomized controlled trial, ignoring missing data results in analyzing a subset of randomized patients within which randomization can no longer be relied upon to balance unmeasured confounders between treatment groups. Moreover, such a subset of patients with complete cases is no longer the intention-to-treat (ITT) population.

There are mounting literature on imputation methods for continuous data following contributions from Donald Rubin, Roderick Little, and many others. Imputation methods for discrete data is an area with growing scientific literature. Hot dec imputation is an approach that is increasingly used to impute continuous and discrete data. In hot dec imputation, missing values for one or more variables of a patient are replaced with observed values from a patient with non-missing data that is similar to the non-respondent with respect to characteristics observed by both cases [2]

In this brief note, we first describe a frequency distribution approach for imputing count data. We then describe a distance-based method for imputation of discrete outcome data using SAS®. The method is based on minimum distance between baseline covariates of those with missing data and those without missing data. The imputation algorithm imputes missing values that are "close to" the observed values, implying that had there been data on those missing, it would have been similar to those non-missing. An illustration is presented.

METHOD

Frequency Distribution Approach for Imputing Discrete Data

Missing count data may be imputed using a frequency distribution approach. Suppose the outcome to be imputed is number of gadolinium enhancing lesion (Gd). To proceed, one could generate the frequency distribution in terms of the numbers of Gd lesions for non-missing subjects. Then one could randomly assign the missing subjects to the categories of the number of lesions proportionate to the percentage of numbers of Gd non-missing subjects.

This uses as a basis for imputation the notion that had patients with missing data produced data, that data would follow the distribution of those with non-missing data. This can be done in three ways: Lump data across treatment group (this is consistent with operating under the null hypotheses); 2) impute for each treatment group using only the non-missing data in each group; 3) impute missing for all groups using only the non-missing data in the placebo group (this is the most conservative as placebo patients could not manifest a treatment effect)

Of note is that as one fills the number of lesions categories based on non-missing for patients with missing, one will have to assign missing for a few patients as having a response or not randomly per the non-missing proportions. For example, if we had 10 missing with percentages 40% lesions and 60% non-lesions based on non-missing, then one randomly selects from the 10 missing patients 4 patients and assign their response a lesion. However, if we had 10 missing with 33% lesions and 67% non-lesions based on non-missing, then one could randomly select 3 patients and assign them lesions. A fourth patient would have to be selected at random and assigned a lesion with a probability of 3%.

A major limitation of the frequency distribution approach is that it ignores the fact that subjects' baseline characteristics may be correlated with outcome. It may be possible to improve the imputation by matching with non-missing subjects in terms of baseline characteristics that are thought to be correlated to development of lesions.

Mahalanobis Distance

The Mahalonobis distance (MD) is a measure of multivariate distance between elements in different units. Denote by X a matrix of covariates. Denote MD between covariates X_i and X_j for two units i and j as:

$$MD(X_i, X_j) = \sqrt{(X_i - X_j)^T S^{-1} (X_i - X_j)} \tag{1}$$

where S and $(X_i - X_j)^T$ are the sample covariance matrix and transpose of $X_i - X_j$, respectively. The matrix X contains covariates which are to be matched between units. MD account for covariance among variables and it also recognizes that variance in each direction could be different.

To avoid possible numerical instability and computational cost associated with inverting variance-covariance matrix S, the Euclidean distance obtained by standardizing covariates can be used. We used procedures available in SAS/ETS software mainly for sake of accessibility and convenience. To obtain an identity covariance matrix of the covariate, the PROC PRINCOMP and PROC SCORE [3] were used; thereafter, FASTCLUS procedure [4] was used to compute Euclidean distance.

Imputing Discrete Data Based on Minimum Distance

Suppose in a data set Z of 12 patients, the non-missing data set NM has 10 patients and the missing data set M has 2 observations. Using data set NM and an appropriate predictive model, one finds baseline characteristics X_i and X_j (say) that are correlated with the outcome to be imputed. Following identification of these baseline characteristics that are correlated with the outcome, one then finds patients in data set NM that are similar to the patients in data set M based on the identified baseline characteristics.

These baseline characteristics are the variables upon which the patients in the two data sets will be matched. Further, suppose that X_i and X_j are variables upon which one wants to match a subset of NM patients to those of M . Using minimum distance, one selects each patient from M and finds the patient in NM that is closest in terms of similarity in baseline characteristics.

We used minimum distance as a similarity measure. Other similarity measures that may also be used are propensity score [5] or genetic algorithm (see [6] for example). However, contextual sensibility of the chosen similarity measure should be carefully evaluated.

The outcome of the selected patient in NM is then used to impute the outcome of patient in M that is closest in baseline characteristics. Figure 1 is a schematic representation of the imputation algorithm.

The steps involved in the imputation are:

1. Separate the original data set Z into two subsets: data set M contains subjects with missing outcome, and data set NM contains subjects with complete data. This can be achieved by the following chunk of code:

```
data NM  M;
set Z;
if missing(gd) then output M;
else if gd^=. then output NM;
run;
/*standardize continuous variables to principal
 component corresponding to age and gbles.
 This adds prin1 and prin2  to the data set*/

proc princomp data=nm std out=outnm
outstat=outstat noprint;
var age gbles;
run;

proc score data=m score=outstat out=outm;
var age gbles;
run;
```

2. sort the data set M in ascending order of the variables to be matched with:

```
proc sort data=outM;
by prin1 prin2;
```

3. Select first subject in data set M and compute the distance between this first subject and every subject in NM

```
proc fastclus maxc=1 replace=none maxiter=0 noprint
  data=outNM
  seed=outM
  out=distance;
  var prin1 prin2;
run;

proc sort data=distance;
by distance;
run;
```

```
data distance1;
set distance(obs=1);
run;
```

4. Remove subject with minimum distance in outNM data set and the selected first subject in outM; Repeat 1 to 3.
5. Repeat 1 to 4 until no observations is left in outM data set (see Figure 1)

Sensitivity Analyses: Imputation under Null, Alternative, and Using Placebo Rate

Irrespective of the statistical sophistication of the approach taken to address missing data, no single approach can overcome the limitation of not having complete data. Hence, sensitivity analysis should follow any chosen method [7]. When statistical analyses aim to compare treatment groups A and B in a randomized clinical trial, three possible (among many) sensitivity analyses using imputation include a) under the null hypotheses, b) under the alternative hypotheses c) using placebo data points.

Implementing distance-based imputation under the null involves matching and imputing disregarding treatment labels. For imputation under the alternative, the imputation is done by treatment group; that is, the imputation is done separately within each treatment group. This ensures data for a treated patient is used to impute outcome of treated patient with similar baseline characteristics. Similarly, data for a placebo patient is used to impute outcome data for a placebo patient with missing outcome.

ILLUSTRATION

A publicly available data set is used to demonstrate imputation of count data using minimum distance because of the confidential nature of Gd lesion data set used for the present application. Moreover, by using a publicly available data set, performance of the method can be judged since results of the original data are readily assessable.

The data set contains attendance information on 314 high school juniors from two schools. The outcome variable is number of days absent from school *daysabs*. The variable *math* is the standardized math score for each student. The variable *prog* indicates the type of instructional program in which the student is enrolled. Predictors of the number of days of absence are the program type and a standardized test in math. Prior analyses of this data set may be found at http://www.ats.ucla.edu/stat/r/dae/nbreg.htm; this site is maintained by the Institute for Digital Research and Education, University of California, Los Angeles. We set approximately 15% of the outcome variable *daysabs* to missing and imputed the missing cases using the method described above. Parameter estimates based on negative binomial model from the original data set, complete cases, and data set with missing data points imputed using minmum distance are presented in Table 1. When more than one approach are used to impute missing data, a measure of performance of an imputation method may be obtain by $\sum(D_i - E_i)^2$, where D_i is the value of the ith actual data point that is declared missing and E_i is it's imputed or estimated value. The summation is over imputed cases.

CONCLUSION

A distance-based method may by used to impute missing discrete outcome data. An example application of this imputation method was used in the imputation of missing data in the optimal dosage design for Cimetidine trial supported by the second author in 1985 (chapter 12 of [8]). The present application is in the imputation of missing Gd lesions in an exploratory statistical data analyses.

As with any missing data imputation approach, sensitivity analyses should be considered as a critical part of imputation of missing count data. Contextual assumption of and interpretation of results from statistical analysis of imputed data should be clearly stated. For instance, does the results imply that the

Data set Z			
SUBJID	X1	X2	Y
001-001	63	2	4
001-002	27	2	4
001-003	20	2	2
001-004	16	3	.
001-005	2	4	3
001-006	71	5	13
001-007	63	3	11
001-008	3	2	7
001-009	51	4	10
001-010	49	3	9
001-011	31	2	.
001-012	22	1	5

Separate Data set *Z* into *NM* and *M*

Data Set *NM*			
SUBJID	X1	X2	Y
001-001	63	2	4
001-002	27	2	4
001-003	20	2	2
001-005	2	4	3
001-006	71	5	13
001-007	63	3	11
001-008	3	2	7
001-009	51	4	10
001-010	49	3	9
001-012	22	1	5

Data set *M*			
SUBJID	X1	X2	Y
001-004	16	3	.
001-011	31	2	.

1) Sort the data set ***M*** in ascending order of variables to be matched
2) Select first subject in ***M*** and compute distance between this subject and all subjects in ***NM***;
3) Select subject in ***NM*** with smallest distance to the first subject in ***M***; remove both subjects and keep them in a data set ***OUT***;
4) Repeat 2 and 3 and add the selected subject to the subject list in ***OUT***.
5) Repeat 2, 3 and 4 until last subject in data set ***M***.

The data set ***OUT*** will contain matched subjects in ***NM*** and M. The number of subjects in data set ***OUT*** will be determined by the number of subjects in ***M***.

Data set OUT								
Subset of NM that matched with M patient					Match M data set with Y imputed based on corresponding matching in NM			
SUBJID	X1_match	X2_match	Y_nm		SUBJID	X1	X2	Y_impute
001-003	20	2	3		001-004	16	3	3
001-002	27	2	4		001-011	31	2	4

Figure 1: A Schematic Representation of the Imputation Algorithm

Table 1: Analysis Of Maximum Likelihood Parameter Estimates

A: Original Data (N = 314; None Missing)				
Parameter	Estimate (SE)	Wald CI	Wald χ^2	P-value
Intercept	2.6153 (0.1964)	2.2304, 3.0001	177.40	<0.0001
Math	-0.0060 (0.0025)	-0.0109, -0.0011	5.71	0.0168
Prog=2	-0.4408 (0.1826)	-0.7986, -0.0829	5.83	0.0158
Prog=3	-0.2787 (0.2020)	-1.6745, -0.8828	40.08	<0.0001
Prog=1	0.0000 (0.0000)	0.0000, 0.0000	0.0000	
Dispersion	0.9683 (0.0995)	0.7916, 1.1844		
B: Subset [(84%; N = 262) of Original Data]				
Intercept	2.4887 (0.2057)	2.0858, 2.8918	146.42	<0.0001
Math	-0.0056 (0.0028)	-0.01111, -0.0002	4.16	0.0413
Prog =2	-3426 (0.1943)	-0.7234, -0.0382	3.11	0.0778
Prog=2	-1.1430 (0.2180)	-1.5703, -0.7156	27.46	<0.0001
Prog=1	0.0000 (0.0000)	0.0000, 0.0000	0.0000	
Dispersion	0.9392 (0.1063)	0.7523, 1.1726		
C: Imputed Data Set [B with 52 (16%) missing imputed)				
Intercept	2.4936 (0.1883)	2.1246, 2.8626	175.40	<0.0001
Math	-0.0056 (0.0025)	-0.0104, -0.0008	5.27	0.0216
Prog=2	-0.3694 (0.1781)	-0.7186, -0.0203	4.30	0.0381
Prog=3	-1.1506 (0.1988)	-1.5402, -0.7611	33.51	<0.0001
Prog=1	0.0000 (0.0000)	0.0000,0.0000	0.0000	
Dispersion	0.9096 (0.0954)	0.7406, 1.1172		

patients whose data were imputed had 100% compliance to treatment? Before implementing any imputation method, mechanisms that generated the missing data should be considered. Modifications to above method that propagates missing data uncertainty are possible.

ACKNOWLEDGEMENT

We thank SAS® technical team, Tom Abernathy, and Rick Wicklin for their support and assistance during development of the macro.

References

[1] D. L. Fairclough, *Design and Analysis of Quality of Life Studies in Clinical Trials*, 2nd ed. Chapman and Hall/CRC, 2010.

[2] R. R. Andridge and R. J. A. Little, "A review of hot deck imputation for survey non-response," *International Statistical Review*, vol. 78, no. 1, pp. 40–64, 2010.

[3] W. W. Feng, Y. Jun, and R. Xu, "A method/macro based on propensity score and mahalanobis distance to reduce bias in treatment comparison a method/macro based on propensity score and mahalanobis distance to reduce bias in treatment comparison in observational study," *pharmasug*, no. Paper PR05, 2006.

[4] SAS/STAT®, *The FASTCLUS PROCEDURE 13.2 User's Guide*, SAS Institute.

[5] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score the central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, pp. 41–55, 1983.

[6] T. K. Tai and K. E. Peace, "Analysis of subgroup data of clinical trials," *Journal of Causal Inference*, vol. 1, no. 2, pp. 193–207, 2012.

[7] G. Molenberghs and M. G. Kenward, *Missing Data in Clinical Studies*. West Sussex PO19 8SQ, England: John Wiley and Sons Ltd, 2007.

[8] K. E. Peace and D.-G. D. Chen, *Clinical Trial Methodology*. Chapman and Hall/CRC, 2010.

CONTACT INFORMATION (AND DISCLAIMER)

Your comments and questions are valued and encouraged. Contact the author at:

Macaulay Okwuokenye, BIOGEN INC, MA, USA

E-mail address: macaulay.okwuokenye@biogen.com