IJRSET Volume 1, Issue 1 Pages: 44-49

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

A SURVEY ON MISSING DATA AND METHODS TO FIND THE MISSING VALUES

[1] P.Logeshwari, [2] Dr. Antony Selvadoss Thanamani
[1] PhD Research Scholar, [2] Associate Professor,
[1, 2] Department of Computer Science,
[1, 2] Nallamuthu Gounder Mahalingam College, Pollachi, India.

ABSTRACT

Missing data plagues almost all surveys, and quite a number of designed experiments. No matter howcarefully an investigator tries to have all questions fully responded to in a survey, or how well designedan experiment is; examples of how this can occur are when a question is unanswered in a survey, or aflood has removed a crop planted close to a river. The problem is, how to deal with missing data, once ithas been deemed impossible to recover the actual missing values. Traditional approaches include case deletion and mean imputation. These are thedefault for the major Statistical packages. In the last decade interest has centered on RegressionImputation, and Imputation of values using the ExpectationMaximization algorithm, both ofwhich will perform Single Imputation. More recently Multiple Imputation has become available, and isnow being included as an option in the mainstream packages.

Keywords: - [Missing Data, Imputation techniques, missing methods, Data Identification]

1. SURVEY ON MISSING DATA

Historical Development Until the 1970s, missing values were handled primarily by editing. Rubin (1976) developed a frame-work of inference from incomplete data that remainsin use today. The formulation of the EM algorithm (Dempster, Laird, & Rubin, 1977) made it feasible tocompute ML estimates in many missing- data problems. Rather than deleting or filling in incompletecases, ML treats the missing data as random variablesto be removed from (i.e., integrated out of) the likelihood function as if they were never sampled. Weelaborate on this point later after introducing the notion of MAR. Many examples of EM were describedby Little and Rubin (1987). Their book also documented the shortcomings of case deletion and singleimputation, arguing for explicit models over informalprocedures. About the same time, Rubin (1987) introduced the idea of MI, in which each missing value isreplaced withm> 1 simulated values prior to analysis. Creation of MIs was facilitated by computer technology and new methods for Bayesian simulation discovered in the late 1980s (Schafer, 1997). ML and MI are now becoming standard because of implementations in free and commercial software

2. MECHANISMS OF MISSING MCAR

The term 'Missing Completely at Random' refers to data where the missingness mechanism does not depend on the variable of interest, or any other variable, which is observed in the dataset. MCAR is bothmissing at random, and observed at random (This means the data was collected randomly, and does notdepend on any other variable in the data set). This very stringent condition is required in order for casedeletion to be valid, and missing data is very rarely MCAR (Rubin, 1976).

MAR

The term 'Missing at Random' is a misnomer, as the missing data is anything but missing at random. Theintuitive meaning of this term is better suited to the term MCAR. What MAR means is missing, butconditional on some other 'X-variable' observed in the data set, although not on the 'Y-variable' of interest(Schafer, 1997).

NMAR

Not Missing at Random, (or informatively missing, as it is often known) occurs when the Missingnessmechanism depends on the actual value of the missing data. This is the most difficult condition to model.Traditional Procedures to find Missing Data Compare the missing and non-missing cases on variables where information is notMissing.

Whatever strategy you follow you may be able to add plausibility to your results (or detect potential biases) by comparing sample members on variables that are not missing. For example, in a panel study, some respondents will not be re-interviewed because they could not be found or else refused to participate.You can compare respondents and non-respondents in terms of demographic characteristics such as race, age, income, etc.

If there are noteworthy differences, you can point them out, e.g. lower -income individuals appear to be underrepresented in the samplesimilarly, and you can compare individuals who answered a question with those who failed to answer. Alternatively, sometimes you may have external information you can draw on, e.g. you know what percentage of the population is female or black, and you can compare your sample's characteristics with the known population characteristics.

Dropping variables

When, for one or a few variables, a substantial proportion of cases lack data, the analyst may simply opt to drop the variables. This is no great loss if the variables had little effect on Y anyway. However, you presumably would not have asked the question if you did not think It was important Still, this is often the best or at least most practical approach. A great deal of missing data for an item might indicate that a question was poorly worded, or perhaps there were problems with collecting the data.

Dropping subjects, i.e. list wise (also called case wise) deletion of missing data. Particularly if the missing data is limited to a small number of the subjects, you may just opt to eliminate those cases from the analysis. That is, if a subject is missing data on any of the variables used in the analysis, it is dropped completely .The remaining cases, however, may not be representative of the population . Even if data is missing on a random basis, a list wise deletion of cases could result in a substantial reduction in sample size, if many cases were missing data on at least one variable.

3. METHODS TO IDENTIFY MISSING DATA

(a). Case Deletion: This can be either list wise (complete case only) or all value (Pairwise-availablecase), the cases are deleted which contain missing data, for the analysis being carried out.

(b). Single Imputation: This can include group means, medians or modes (depending on the data), Regression Imputation, Stochastic Regression Imputation (deterministic regression imputation with an added random error component), or EM Imputation (this uses the Expectation-Maximization algorithm to predict the missing value), or hot deck imputation, or last value carried forward for longitudinal data, and a variety of other methods (Scheffer, 2000). End users Very often demand a single complete data set.

(c). Multiple Imputations: Frequentist MI. This returns m complete datasheets by imputing m times. This can be based on propensity scoring, if imputation model fails to converge. Bayesian MIuses MCMC algorithm with a non-informative prior to predict the posterior distribution fromwhich random draws are made, producing m individual datasheets. Successful multipleimputation may be shunned by an end-user, as the concept of more than one datasheet for aparticular survey is daunting to non-statisticians. However, multiple imputation is always betterthan case deletion, or single ad-hoc methods

(d) Mean imputation within classes (MC). Thismethod divides the total sample into imputationclasses according to values on the auxiliaryvariables. The classes may be defined as all the cells in thecross-tabulation of the (categorized) auxiliaryvariables, but this symmetry is not essential;instead, some auxiliary variables may be used forone part of the sample while others are used foranother part, or groups of cells may be combined. If all the cells in the cross-tabulation are used,the linear function can be expressed as a modelwith the main effects and all levels ofinteraction for the auxiliary variables. In general, the model can be represented by Ymi = bro +Y~brjzji, where the zji are dummy variables, zji = I if the i-th non respondent is inclass j, zji = 0 otherwise (j = 1,2,...,(H- I)).Since emi = 0, the method is a deterministic one.

(e) Random imputation within classes (RC). Thismethod corresponds to the random overall methodexcept that it is applied within imputationclasses. Each non respondent is assigned the yvalueof a respondent randomly selected from thesame imputation class. The method is thestochastic equivalent of the mean within class method, respondent residual selected at random withinimputation class j in which non respondent is located.

(f) Hot-deck imputation. The term hot-deckimputation has a variety of meanings, but refershere to the sequential type of procedure used bythe Bureau of the Census with the labor force items in the Current Population Survey (CPS)(Brooks and Bailar, 1978). This is sometimesknown as the traditional hot-deck procedure. Theprocedure begins with the specification ofimputation classes, and for each class theassignment of a single value for the y-variable to provide a starting point for the process. Thesestarting values may, for instance, be obtained bytaking a respondent value for each class or arepresentative value such as the class mean from aprevious round of the survey. The records of thecurrent survey are then treated sequentially. Ifa record has a response for the y-variable; thatvalue replaces the value previously stored for itsimputation class. If the record has a missingresponse, it is assigned the value currentlystored for its imputation class. A majorattraction of this procedure is its computingeconomy; since all imputations are made from asingle pass through the data file.The hot-deck method is similar to the randomwithin class method in which donors are selectedby unrestricted sampling (i.e. SRS withreplacement). If the order of the records in thedata file were random, the two methods would beequivalent, apart from the start-up process. The sequential hot-deck procedure generally benefitsfrom the non-random order of the data file, sinceuse of the preceding donor in the imputation classyields an additional degree of matching which isadvantageous if the file order creates positiveautocorrelation. This benefit is unlikely to besubstantial, however, when the imputation classesare small and spread throughout the file - as isoften the case.A disadvantage of the hot-deck method is thatit may easily give rise to multiple use of donors,a feature which leads to a loss of precision forthe survey estimators. This occurs when within agiven imputation class a record with a missingresponse is followed by one or more records withmissing responses; all these records are thenassigned the value from the last respondent in theclass s. The random within class method withunrestricted sampling of donors shares thisdisadvantage. With the random within classmethod, however, the multiple use of donors may beminimized by sampling donors without replacement.

It is impossible to develop a model-freetheoretical evaluation for the hot-deck methodbecause of its dependence on the order of the fileand its lack of a probability mechanism. For thisreason, it will not be examined in the subsequentsections; the results for the random within classmethod with unrestricted sampling should, however,provide a reasonable guide to its performance.

Useful discussions of the hot-deck procedure areprovided by Bailar, Bailey and Corby (1978),Bailar and Bailar (1978, 1979), Ford (1980), Ohand Scheuren (1980), Oh, Scheuren and Nisselson(1980) and I. Sande (1979a,b).

(g) Flexible matching imputation. The termflexible matching imputation is used here for themodified hot-deck procedure that has been usedsince 1976 for the CPS March Income Supplement.The procedure sorts respondents and nonrespondentsin t o a large number of imputation classes,constructed from a detailed categorization of asizeable set of auxiliary variables. Nonrespondents are then matched with respondentson a hierarchical basis, in the sense that if anon respondent cannot be matched with a respondentin the initial imputation class, classes arecollapsed and the match is made at a lower level.Three levels are used with the March IncomeSupplement, the lowest level being such thatamatch can always be made. The procedure enablescloser matches to be secured for manynonrespondents than does the traditional hot-deckprocedure. It also avoids the multiple use ofrespondents in classes where the number ofnonrespondents does not exceed the number ofrespondents. Further details on theimplementation and evaluation of the procedure aregiven by Coder (1978) and Welniak and Coder(1980).

(h) Predicted regression imputation (PR). Thismethod uses respondent data to regress y on theauxiliary variables. Missing y-values are thenimputed as the predicted values from the regression equation, Ymi = bro + Y brjzji. This isa deterministic method with emi = O. Theauxiliary variables may be quantitative or qualitative, the latter being incorporated by Means of dummy variables. If the y-variable is qualitative, log-linear or logistic models may beused. As in any regression analysis, specific interaction terms may be included in theregression equation, and transformations of thevariables may be useful.

A special case of the regression model is the ratio model Ymi = brzi with a single auxiliaryvariable and an intercept of zero (Ford, Klewenoand Tortora, 1980). This model may be used inpane i surveys with z representing the samevariable as y measured on the previous wave.

(i) Random regression imputation (RR) .This method is the stochastic version of the predictedregression method: the imputed values are thepredicted values from the regression equation plus residual terms emi. Depending on the assumptionsmade, the residuals can be determined in various way s, including.

(i) If the residuals are assumed to behomoscedastic and normally distributed, a residualcan be chosen at random from a normal distributionwith zero mean and variance equal to the residualvariance from the regression.

(ii) If the residuals are assumed to come from thesame, unspecified distribution, they can be chosenal random from the respondents" residuals.

(iii) As a protection against non-linearity andnon-additivity in the regression model, theresiduals may be taken from respondents withsimilar values on the auxiliary variables. If thedonor respondent has the identical set of z valuesas the nonrespondent, the procedure reduces toa s s i g n i n g t h e r e s p ondent" s y-value to thenonrespondent. This point demonstrates the closerelationship between this procedure and the randomwithin class method.Applications of regression and categorical datamodels for imputation are described by Schieber(1978), Herzog and Lancaster (1980) and Herzog(1980).

(j) Distance function matching. This method assigns the y-value of the nearest respondent toeach nonrespondent, with "nearest" defined by adistance function of the auxiliary variables. Themethod is primarily concerned with quantitativevariables; however, qualitative variables may beincluded either by using the distance function approach within imputation classes formed byqualitative auxiliary variables or byincorporating these variables into the distancefunction. With a single auxiliary variable, thesample may be ordered by the variable, and thenearest respondent (donor) to each non respondent is taken where "nearest" may be defined as theminimum absolute difference between the non respondent" s and donor's values in theauxiliary variable or in some transformation ofthe auxiliary variable. When several auxiliaryvariables are used, the issue of transformationsbecomes more critical; one approach is totransform all auxiliary variables to their ranks.

It can beconstructed to reduce the multiple use of donors. For instance, distance may be defined as D(I + pd)where D is the basic distance, d is the number oftimes the donor has already been used and p is apenalty for each usage (Colledge et al., 1978).

A variant of this method assigns thenonrespondent the average value of neighboringrespondents, for instance the average value of thetwo adjacent respondents (Ford, 1976). As withother averaging procedures, this procedure suffersthe disadvantage of distorting distributions.

k) Deductive imputation. This imputationmethoddepends on some redundancy in the data so that amissing response can be educed from the auxiliaryinformation, i.e. ymi = f(zi) exactly. Forexample, if a record should contain a series ofamounts and their total but one of the amounts ismissing, the missing value can be deduced bysubtraction. The method can be extended to situations where the deduced value is highly likely to be the correct value or at least close to it; for instance, in a panel survey with avariable that remains almost constant over time, amissing response on one wave of the panel may beassigned the record's value for the item on thepreceding or succeeding wave.

(l) Mean imputation overall (MO). This methodassigns the overall respondent mean, Yr, to all missing responses. It is the deterministicdegenerate form of the linear function with no auxiliary variables **(m) Random imputation overall (RO).** This method assigns each no respondent the y-value of arespondent selected at random from the totalrespondent sample. The method is the stochasticdegenerate form of the linear function with noauxiliary variables, Ymi = Yr + emi, withemi = Yrk-Yr, which reduces to Ymi = yrk. Givenan epsem sample initial!y, the subsample o frespondents to act as donors can be selected byanyepsem sampling scheme (e. g. unrestri c t e dsampling, SRS, proportionate stratified sampling,or systematic sampling).

CONCLUSIONS

A major attraction of imputation is that itgenerates a complete data set that may be readilyused for many different forms of analysis. As thepreceding sections have shown, however, caution is needed in analyzing a data set that includesimputed values. In the case of univariate analyses, deterministic imputation methods servewell for estimating means and totals, but theydistort the distributional properties of thevariable; stochastic methods are less efficientfor estimating means and totals but they preservethe variability in the respondent data. Allmethods are likely to attenuate the covariance'sbetween the variable subject to imputation andother variables, except for those other variablesthat are used as auxiliary variables in theimputation scheme. In consequence, when a dataset contains imputed values, special care isneeded in studying the interrelationships betweenvariables, whether the interrelationships a r eexamined in terms of cross-tabulations, regressionanalyses or other forms of multivariate analysis. Alternative ways of handling missing surveydata include dropping cases with missing values onthe relevant variables from the analysis, directestimation of the population parameters from amodeling approach, and weighting adjustments. Dropping cases with missing values is a widelyused procedure, sometimes adopted on the groundsthat it avoids assumptions required in procedures Which attempt to compensate for missing data.

REFERENCES

[1] Little, R.J.A. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83, 1198-1202.

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

[3] Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 581-592.

[4] Rubin, D.B. and Schenker, N. (1991). Multiple imputation in health-care data bases: An overview an some applications. Statistics in Medicine, 10, 585-598.

[5] Schafer, J.L. (2000). NORM. Version 2.03. http://www.stat.psu.edu/~jls/.

[6] Schafer, J.L. and Graham, J.W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7, 147-177.

[7] van Buuren, S., Brands, J.P.L., Groothuis-Oudshoorn, C.G.M., and Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation, 76, 1049-1064.

[8] van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research, 16, 219-242.

[9] van Ginkel, J.R., van der Ark, L.A., and Sijtsma, K. (2007). Multiple imputation of test and questionnaire data and influence on psychometric results. Multivariate Behavioral Research, 42, 387-414.