Diagnostics for Multiple Imputations

Trivellore Raghunathan and Irina Bondarenko [1]

Abstract

Multiple imputation technique is becoming a popular method for analyzing data with missing values. Several methods have been proposed for creating multiple imputations and most of these methods assume that the data are missing at random (MAR). However, limited diagnostic tools are available to check whether the imputations created by these methods are reasonable. This article develops a set of diagnostic tools based on certain conditional distributions of the observed and imputed values. These conditional distributions should be similar if the assumed model for creating multiple imputations is a good fit. The tools are formulated in terms of numerical summaries and graphical displays and could be easily implemented using the standard complete data software packages. For implementing these methods the exact nature of the model used by the imputer is not needed. The method is illustrated using a data set with large number of variables of different types with varying amount of missing values.

Key words: Congeniality, Diagnostics, Missing at Random, Propensity score matching, Residuals

1 Introduction

Many data analyses involve incomplete data. Several methods have been proposed for analyzing incomplete data for particular statistical models such as multivariate normal, loglinear models etc (see Little and Rubin (2003) for a review) and software for implementing them

[1]Trivellore Raghunathan is Professor of Biostatistics in the School of Public Health and Research Professor at the Institute for Social Research. Irina Bondarenko is Statistician Senior in the Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor MI 48109. Research was supported by NIH grants P01-HD045753 and R24-HD047861

have become available. However, multiple imputation approach (Rubin (1978, 1987)) is more convenient because it exploits the complete data software to perform the analysis and hence, is more versatile in terms of handling a variety of statistical models. In the multiple impuation approach, the missing set of values are replaced by several plausible sets of values. Each plausible value together with the observed set of values forms a completed data. Each completed data set is analyzed using the standard complete data software. The completed data point estimates of parameters, their the standard errors and other statistics are combined to form a single inference (Rubin and Schenker (1986), Li et al (1991a), Li et al (1991b), Meng and Rubin (1994) and Barnard and Rubin (1999)). Software for creating multiple imputations and combining completed data analysis is also becoming more available.

Like any other approach for handling missing data, imputation based approaches also make a set of assumptions and use implicit (such as Hot-deck, Predictive mean matching, Last observation carried forward etc (Little and Rubin (2002)) or explicit (such as multivariate normal, linear regression etc) model assumptions (Schafer (1997)). In addition, most imputation models assume that the data are missing at random [MAR] (Rubin (1976)). Random draws from an approximate predictive distribution of the missing values under the specific model are then used as imputations. Multiple imputation involves repeated independent drawing of values from this predictive distribution.

Many practical situations involve complex data structures with missing values in several types of variables (such as continuous, ordinal, nominal, count, semi-continuous etc) and skip patterns (for example, some variables are not applicable to a particular group of subjects) and restrictions (such as years of smoking which are necessarily bounded by the age of the person). For such complex structures a sequential regression multivariate imputation approach has been proposed (Kennickel (1992), Van Buuren and Oudshoom (2000) and Raghunathan et al (2001)). These are Gibbs sampling type iterative procedures in which the missing values in each variable are imputed conditional on all other variables using appropriate regression models. The imputations are draws from the corresponding approximate posterior predictive distribution of the missing values.

Though several methods for creating imputations are available, diagnostic tools to check the validity of imputed values under the stated assumptions are not well developed. As

suggested in Abayomi, Gelman and Levy (2005), it is possible to check the imputation model fit using the observed data and appropriately fine tune the models (for example, transforming the outcome to achieve normality, if the residual diagnostics indicate lack of normality). However, even if the model fits the data well, randomly generated values may not be reasonable. For example, if the log-normal model is deemed to be appropriate for imputing missing income values based on regression diagnostics, the draws from the posterior predictive distribution under this model when back transformed to the original scale could be untenably large compared to the observed values (He and Raghunathan (2006)).

Another possibility is to compare the histograms, descriptive statistics or the frequency distribution of the observed and imputed values for each variable. Such print outs are commonly produced by software for creating imputations (see Royston (2004), Raghunathan et al (1997)). One would expect much agreement between the observed and imputed marginal distributions only under missing completely at random (MCAR) mechanism.

Consider as an example, Alameda County Study (Hochstim (1970), Berkman and Breslow (1983)) a large probability sample cohort of approximately 7000 subjects selected in 1965 from the Alameda County, California. We have selected approximately 150 variables and multiply imputed the missing values using the sequential regression approach, implemented in IVEWARE (www.isr.umich.edu/src/smp/ive). We created $M = 10$ imputations.

One of the variables with missing values, *Havewage,* was a response to the question "Did you receive income from wages and salary during the past year?" Approximately 4219 (63%) subjects answered "Yes" and 2513 (37%) subjects answered "No" with 196 people electing not to answer that question. The missing data percentage is rather small but nevertheless quite useful to illsutrate the pitfalls of comparing the marginal distribution. In the first imputed data set, 60 (31%) were imputed to "Yes" and 136 (69%) were imputed to "No". The observed and imputed percentages are almost reversed. The same pattern was observed in the remaining 9 imputed data sets. One may suspect that there is something wrong with the imputation model based on this comparison. It is unlikely (in fact quite rare) under the MCAR assumption but it is not clear whether this is suspicious under the MAR assumption.

As the second example, consider a continuous variable, the number of years of smoking for the current or past smokers. This is relevant only for the current/past smokers and some

current/past smokers did not answer this question. Figure 1 gives a pair of kernel densities for the observed and imputed values for the current/past smokers based on the first imputed data set. Similar pattern was observed in other imputed data sets. It appears that the observed and imputed values are different. In fact, both Kolmogrov-Smirnov and Kuiper tests for comparing the two distributions rejects equality for every imputed data set. Does this mean that there is something wrong with the imputation model? The answer is yes, if the assumed mechanism is MCAR. But it is not clear what one would expect under MAR.

In many research groups, multiple imputation may have been performed for a large set of variables. Such strategies are operationally more efficient and may also be more statistically efficient, if the model used by the imputer include sveral predictors of the variables with missing values. However, an analyst may be interested in fitting a model involving a subset of variables that may or may not be congenial with the imputer model (Meng, 1995, 2002). Since the analyst's inferences are conditional on the correctness of his/her model, it is prudent to know whether the imputation created by the imputer for the subset of variables is reasonable under the missing at random assumption for the specific analyst model. If it is not, then analyst may want to perform imputation under his/her model. Model diagnostics can be used only if the model used by the imputer is available. If the imputer and analyst are different people or in different organizations then such information may not be available. We need a set of tools to compare the distributions of the observed and imputed values that the user could use in the context of his/her analysis model without the exact knowledge of the model used by the imputer.

2 Proposed Diagnostic Method

Suppose that the survey data set has n observations and p variables, $Y_v, v = 1, 2, \ldots, p$. For simplicity assume that the survey data is a simple random sample. For complex survey designs, we can use the design variables as predictors while imputing the missing values, a strategy described and evaluated in Rieter, Raghunathan and Kinney (2006).

For the variable Y_v with missing values assume, without loss of any generality, that the observed values for subject s are $y_{obs,v} = \{y_{sv}, s = 1, 2, \ldots, n_v\}$ and $y_{mis,v} = \{y_{sv}, s =$

$n_v + 1, n_v + 2, \ldots, n\}$ is the set of missing values. Let $R_{sv} = 1, s = 1, 2, \ldots, n_v$ and $R_{sv} = 0, s = n_v + 1, n_v + 2, \ldots, n$ be the response indicator. Let $Y_{obs,-v}$ denote the observed set of values of $Y_i, i = 1, 2, \ldots, v-1, v+1, \ldots, p$. Let $e_{obs,-v} = Pr(R_v = 1|Y_{obs,-v})$ be the true response propensity for variable Y_v as a function of the observed data. The propensity score is an efficient summary of the covariates $Y_{obs,-v}$ and could be used compare the outcome across the two treatment groups $R_v = 1$ and $R_v = 0$ (Rosenbaum and Rubin (1983)) adjusted for $Y_{obs,-v}$. Thus, under MAR mechanism we expect that, conditional on $e_{obs,-v}$, the distributions of $y_{obs,v}$ and $y_{mis,v}$ to be similar.

For $l = 1, 2, \ldots, M$, let $y^{(l)}_{mis,v} = \{y^{(l)}_{sv}, s = n_v + 1, n_v + 2, \ldots, n\}$ denote the imputed set of values for missing set $y_{mis,v}$. If the imputations are reasonable under the missing at random assumption, then the observed set $y_{obs,v}$ and each imputed set $y^{(l)}_{mis,v}$ should have similar distributions conditional on the propensity score $e_{obs,-v}$. For example, we could create quintiles based on the true response propensity and check the equality of the distributions of the observed and missing values within each quintile. For discrete variables we could compare the frequency distribution of the observed and imputed values. For continuous variables we could compare the histograms or descriptive statistics within each quintile. Alternatively, we could regress the observed and imputed values of Y_v on the true propensity score and compare the distribution of the residuals between the respondents and nonrespondents.

Obviously, the true propensity score $e_{obs,-v}$ is unobserved and needs to be estimated for each subject. It may appear to be a daunting task, but multiply imputed data sets can be used to estimate this propensity score.

The true propensity score $e_{obs,-v}$ can be written as

$$e_{obs,-v} = \int Pr(R = 1|Y_{obs,-v}, Y_{mis,-v}) pr(Y_{mis,-v}|Y_{obs,-v}) dY_{mis,-v}$$

as an average of the true complete data response propensity score, $e_{-v} = Pr(R = 1|Y_{obs,-v}, Y_{mis,-v})$ with respect to the posterior predictive distribution of $Y_{mis,-v}$ conditional on $Y_{obs,-v}$.

Let $Y^{(l)}_{mis,-v}, l = 1, 2, \ldots, M$ denote the M sets of imputed values for the missing values in all the variables except Y_v.Under the correct imputation model and for sufficiently large M, the above true response propensity score can be approximated by

$$e_{obs,-v} = \sum_{l}^{M} e^{(l)}_{-v}/M$$

where $e_{-v}^{(l)} = Pr(R = 1|Y_{obs,-v}, Y_{mis,-v}^{(l)})$ is the completed data response propensity score.

What remains is to estimate the completed data response propensity from each completed data set for each subject and then take the average across the M response propensities. It is fairly easy to estimate the completed-data response propensity model by using, for example, a logistic or probit regression model with R_v as the outcome variable and the completed data $Y_i, i = 1, 2, \ldots, v-1, v+1, \ldots, p$ as predictors. That is, for sufficiently large M, we can approximate

$$\widehat{e}_{obs,-v} = \widehat{Pr}(R_v = 1|Y_{obs,-v}) = \sum_{l=1}^{M} \widehat{Pr}(R_v = 1|Y_{obs,-v}, Y_{mis,-v}^{(l)})/M,$$

In estimating the completed data response propensities, one could include interaction terms in addition to the main effect terms; Use goodness of fit statisitics proposed, for example, in Hosmer and Lemeshaw (1989) to fit completed data response propensity model. One could also use nonparametric regression models such as generalized additive models to estimate the observed data propensity scores.

3 Examples Revisited

Revisiting the imputation of variable *Havewage*, we estimated the propensity scores from each completed data set and averaged them across the $M = 10$ data sets. We then classified the imputed and observed values into four classes based on the top three quintiles and the two bottom quintiles combined into a single group. The two bottom quintiles were combined due to small number of missing values. The number of imputed values in the three groups was 130, 37, 20 and 9 respectively. The first imputed data and the observed data proportion of subjects reporting receiving income from wages and salaries were (20.4, 21.5), (38.6, 37.8) (70.0, 71.8) and (44.4,89.1). The last cell being quite small is unstable. Similar pattern was observed in the remaining 9 imputed sets. When averaged across all 10 imputations, the imputed and observed proportions in the four cells were (19.6, 21.5), (41.9,37.8), (73.0,71.8) and (68.0,89.1). Even though marginally, the imputed and observed proportions were very different but conditional on the estimated propensity score, the observed and imputed proportions are closer.

For the continuous variable, *yearssmoked*, Figure 2 plots the observed and imputed values against propensity scores. Assuming MAR and correct imputation model, we expect the distributions of the observed and imputed values to be similar conditional on the propensity score. That is, for any given value of the propensity score, there should be no discernable differences between the subjects with imputed and observed values. Again, this example illustrates that marginal distribution of the imputed and observed subjects may be different but it could be erroneous to conclude that imputed values are unreasonable. An alternative graphical display is given in Figure 3, which compares the kernel densities of residuals after regressing the observed and imputed values on the propensity scores. Under MAR assumption, the two kernel densities should be overlapping as they do in Figure 3. To summarize these results numerically, we repeated Kolmogorov-Smirnov and Kuper tests, comparing residuals. Both tests failed to reject the equality of the conditional distributions of observed and imputed values for all the imputed datasets.

4 Discussion

In the context of multiple imputations, we have proposed a simple method for checking whether the imputed values are reasonable under MAR without knowing exactly the model used by the imputer. The approach involves estimating the observed data response propensity score based on all the variables except the variable under consideration and then comparing the distributions of observed and imputed values conditional on the estimated propensity score. The propensity score can be estimated by averaging the completed data estimated propensity scores. This method can be implemented using standard software packages as it involves repeatedly applying logistic or probit regression models to estimate the propensity score.

Such tools can be useful to both imputers and analysts. An imputer can use these diagnostics to check reasonableness of the imputed values under MAR before disseminating it to various analysts. An analyst working with a subset of variables can use these tools to check whether the imputations are compatible with the MAR assumption within this subset. If it is not, then the analyst has the option of re-imputing the missing values within

the subset assuming MAR (that is the analyst model is considered as the gold standard) or may conclude that MAR assumption conditional on the subset is not reasonable. At this point the analyst may use the original imputed values in the analysis (that is consider the imputer model as the gold standard) or perform sensitivity analysis under nonignorable missing data mechansims. On the other hand, if the diagnostic tools applied on a subset do indicate that the imputed values are reasonable, then the inferences using the imputed values will not be that different from the ones obtained by the analyst working fresh with the incomplete data on the subset.

The proposed methodology can be extended in many ways. For example, nonparametric regression approaches can be used to estimate the observed data propensity scores. Generalized additive models can be used to estimate the completed data propensity scores and then averaged across the M completed data sets. Such modifications can be easily implemented given that software for performing such regressions is readily available in standard software packages.

External data sets can be used in the propensity score matching. Suppose that an external data set is available with several common set variables. Each completed data can be appended to this external data and then estimate the propensity score. The average of the appended completed data score can used to match and compare the distributions of observed values in the internal and external data sets with the imputed sets of values. For example, in an economic survey, the Current Population Survey (CPS) can be used as an external information to check both the observed and imputed income values conditional on a set of common covariate values.

Like any other diagnostic tools, comparisons of the observed and imputed values can be used to check for plausibility. These are not that different from residual diagnostic plots and summaries used in the standard regression analysis in that the assessments are subjective. These are not formal tests of hypotheses. For categorical variables with missing values, we can compare the imputed and observed response category percentages within the propensity score classes. This comparison can be made formal by computing chisquare type statistics and comparing it to some reference distribution. These may be useful but likely to noninformative when the sample size is large because any difference becomes statistically

detectable.

Similarly, for continuous outcomes we can use formal test statistics to compare the distributions of observed and imputed values within the propensity score class or the distributions of the residuals after regressing the propensity score, $e_{obs,-v}$ on Y_v. Again such tests may achieve statistical significance because of large sample size.

In this article, we have checked the conditional distributions of the observed and imputed values on a variable by variable basis. It is possible to extend the same strategy to check multivariate conditional distributions. Specifically, suppose that we are interested in checking the joint distributions of the observed and imputed values of (Y_1, Y_2). Let $e_{obs,-(1,2)}$ be the propensity score comparing subjects with both observed versus those with at least one imputed variable conditional on all the variables except Y_1 and Y_2. We can use graphical displays such as P-P plots to compare bivariate residuals or the conditional distributions given the propensity scores for the four groups:(1) both Y_1 and Y_2 observed; (2) Y_1 imputed, Y_2 observed; (3) Y_1 observed, Y_2 imputed; and (4) both imputed.

References

Abayomi, K., Gelman, A., and Levy, M. (2005). Diagnostics for Multivariate Imputations. Unpublished manuscript. Department of Statistics, Columbia University, New York.

Barnard J., Rubin, D.B. (1999). Small-sample degrees of freedom with multiple imputation. Biometrika 86:94955

Berkman L.F.and Breslow L. (1983). *Health and ways of living: the Alameda County Study.* New York, NY: Oxford University Press.

He, Y. and Raghunathan, T. E. (2006). Tukey's gh distribution for multiple imputation. *American Statistician*, 60, 251-256.

Hochstim J.R.(1970). Health and ways of living?the Alameda County, California, population laboratory. In: Kessler IJ, Levin ML, eds. The community as an epidemiologic laboratory. Baltimore, MD: Johns Hopkins University Press, 149-76.

Hosmer, D. W., and Lemeshow,S. (1989). *Applied logistic regression.* New York: Wiley.

Kennickell, A. B. (1991). Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation. *Proc. Sec. Surv. Res. Meth., Am. Statist. Assoc.,* 1-10.

Li K.H., Meng X.L., Raghunathan T.E., Rubin, D.B. (1991a). Significance levels from repeated p-values with multiply imputed data. Statist. Sinica 1:6592

Li K.H., Raghunathan T.E., Rubin, D.B. (1991b). Large sample significance levels from multiply-imputed data using momentbased statistics and an F-reference distribution. J. Am. Statist. Assoc. 86:106573

Little, R. J. A.,and Rubin, D. B., (2002). *Statistical Analysis with Missing Data.* New York:Wiley.

Meng, X.L., and Rubin, D.B. (1992), Performing Likelihood Ratio Tests With Multiply Imputed Data Sets, Biometrika, 79, 103-111.

Meng, X.L. (1995). Multiple imputation with uncongenial sources of input (with discussion), *Statist. Sci.*, 10, 538-573.

Meng, X. L. (2002). A congenial overview and investigation of multiple imputation inferences under uncongeniality, Chapter 23, in *Survey Nonresponse* (R. Groves, D. Dillman, J. Eltinge and R. Little, eds.), New York: Wiley.

Raghunathan, T. E., Lepkowski, J. E., Van Hoewyk, J. and Solenberger, . (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*,27, 85-95.

Raghunathan, T. E., Van Hoewyk, J. and Solenberger, P. (1997). IVEWARE: Imputation and Variance Estimation Software. http://www.isr.umich.edu/src/smp/ive.

Reiter, J. P., Raghunathan, T. E. and Kinney, S.K. (2006). The Importance of Modeling the Sampling Design in Multiple Imputation for Missing Data, To appear in *Survey Methodology*.

Rosenbaum,P.R. and Rubin, D.B. (1983). The central role of the propensity score in the observational studies for causal effects. *Biometrika*, 70, 41-55.

Royston, P. (2004). Multiple imputation of missing values. Stata Journal 4: 227241.

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

Rubin, D. B. and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

Rubin, D. B. (1987). **Multiple Imputation for Nonresponse in Surveys**. New York: Wiley.

Rubin, D.B.(1996). Multiple imputation after 18+ years (with discussion). J. Am. Statist. Assoc. 91:47389

Schafer, J.L. (1997), **Analysis of Incomplete Multivariate Data**, London: Chapman and Hall.

Van Buuren, S. and Oudshoom, C. G. M.(2000). MICE: Multivariate imputation by chained equations. web.inter.nl.net/users/S.van.Buuren/mi/

Figure 1: Kernel density estimates for the marginal distributions of the observed and imputed values.

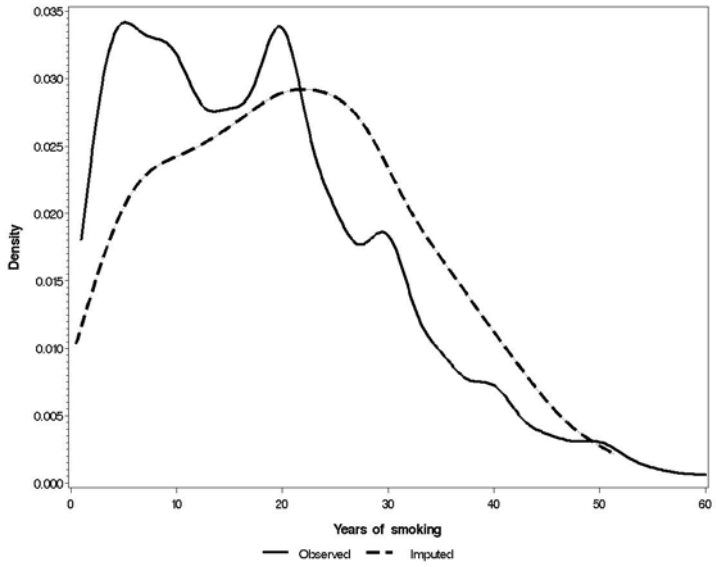

Figure 2: Scatter plot of the observed and imputed values against the propensity scores

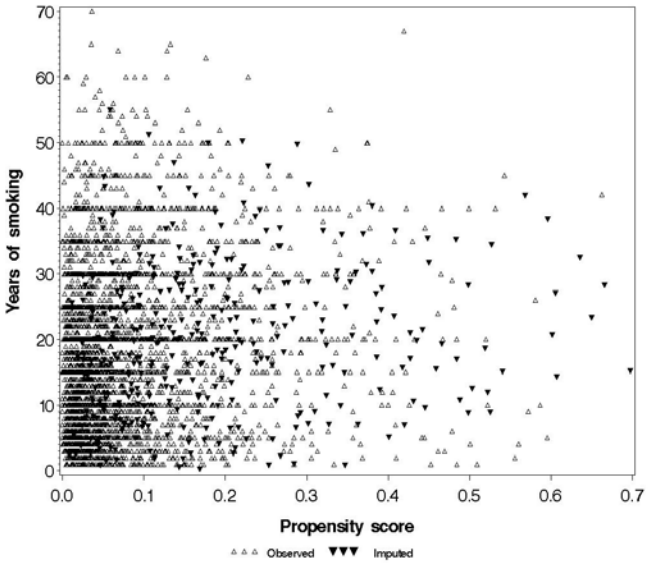

Figure 3: Kernel density estimates of the distributions of the residuals from the regression of the observe/imputed values on the propensity scores.