Adaptive sequential regression imputation methods using machine learning techniques

Trivellore Raghunathan
University of Michigan
Ann Arbor, Michigan, USA

The digital technology, though more than fifty years old, is increasingly yielding data which can be used as a part of statistical infrastructure. Advances in computational resources such as processing powers, storage, development of algorithms and other computational infrastructure have made many administrative records, transactions, electronic health records etc useable by statisticians in combination with the traditional data sources such as probability sample based surveys, and designed data collection activities. The challenge of combining information from these multiple data sources needs to be met by a principled approach for drawing inferences that borrows strength from the two paradigms for drawing statistical inferences: (1) A Frequentist or repeated sampling perspective and (2) the Bayesian framework. The goal of this paper is to outline a set of issues, propose methodology and evaluation strategies in the context of using multiple imputation as a method for constructing inferences from the assemblage data sources.

To motivate the problem, consider a simple context with three data sources: A survey or designed data collection activity (say, activity, A) producing a set of variables (X, Y), and the two organic or non-designed data collection activities (say, activities B and C) producing the sets of variables, (X, Z) and (Y, Z), respectively. Obviously, for a variety of reasons, there may be some missing values in all there sets of variables X, Y and Z. In addition, when these three data sources are appended as single data set, the non-measured sets Z in data A, Y in Data B and X in Data C are also missing, but missing “by assembly”. The goal then is to construct inferences from the assembled data set and validate the inferences, internally, and, possibly, externally, using some other information.

The situation given above can be generalized to more than three data sources and more complex patterns of missing data. Though, it may be unlikely, let U be a collection of variables available in all data sources. However, for simplicity, continue with three data source example to illustrate the methodological development and evaluation strategies. Suppose that n_A, n_B and n_C, respectively, be the number of subjects in the three data sources. Also, let p_X, p_Y and p_Z denote the numbers of variables in each set, X, Y and Z, respectively. When the data set is appended, the full or potential complete data set, D, has $n = n_A + n_B + n_C$ subjects (rows) and $p = p_X + p_Y + p_Z$ variables (columns). Let R be $n \times p$ matrix of response or observed indicators, taking the value 1, if the corresponding entry in the full data set is observed and 0, otherwise.

With a slight abuse of notation, let X_{obs} denote all the values of variables in the set X, across all the subjects in all data sources. Similarly for Y_{obs} and Z_{obs}. Let X_{mis}, Y_{mis} and Z_{mis} denote all the values that are missing in these three sets of variables across all the data sources. Let $D_{obs} = \{X_{obs}, Y_{obs}, Z_{obs}\}$, $D_{mis} = \{X_{mis}, Y_{mis}, Z_{mis}\}$ and $D = \{D_{obs}, D_{mis}\}$.

A Bayesian statistical analyst (See Gelman et al (2004)) conceptualizes the the object of inference as a function of an unknown parameter (possibly a vector), θ, in a potential full data generation model, $f(X, Y, Z|\theta)$. Let $\pi(\theta)$ denote the prior density (or mass) function for θ. Suppose that the full data set is observed and $D = d$ is a particular realization of values in D, the Bayes theorem is used to construct the posterior density,

$$p(\theta|d) = \frac{L(\theta|d)\pi(\theta)}{\int L(\theta|d)\pi(\theta)d\theta},$$

where $L(\theta|d)$ is the likelihood function (proportional to the joint distribution of all n values of p-dimensional vector, evaluated at the realized value of the data d). The posterior density given above is the crux of Bayesian inferences about θ and is constructed through summaries of the posterior distribution such as central tendency, spread, quantiles etc. Much of the modern Bayesian analysis involves drawing values of θ from its posterior distribution and then using the draws to study the features of the distribution.

Obviously, the above method cannot be implemented because not all values in d are known. Only the realized values of $D_{obs} = d_{obs}$ are available and D_{mis} is not known. The inference needs to be extended into inferring about D_{mis} and θ. Additional assumption is needed about the nature of missingness in D. One plausible assumption is that the data are missing at random (MAR) (Rubin (1976)), in a sense that the missing values are “unbiasedly” predictable based on the observed data, or any other additional external information, E . Suppose that a predictive model, $Pr(D_{mis}|d_{obs}, E)$ can be constructed, then this assumption implies that the differences between any randomly drawn value from the predictive distribution, say, D^*_{mis}, and the true or actual unobserved value D_{mis} is a collection of random noises. Note that this assumption is not empirically verifiable as the actual values of D_{mis} are not known. It is critical, therefore, to make sure that d_{obs} (and, possibly, E) are adequate to construct a predictive distribution for the missing values and satisfy this assumption. The MAR assumption may also stated in an equivalent form in terms of,

$$Pr(R|d_{obs}, D_{mis}, E) = Pr(R|d_{obs}, E),$$

a response propensity model indicating which values in D are observed or missing.

Under the stated assumption about the mechanism for the missing data, a modified Bayesian analysis involves constructing,

$$p(\theta|d_{obs}, E) = \frac{\int L(\theta|d_{obs}, D_{mis})\pi(\theta)f(D_{mis}|d_{obs}, E)dD_{mis}}{\int\int L(\theta|d_{obs}, D_{mis})\pi(\theta)f(D_{mis}|d_{obs}, E)dD_{mis}d\theta}$$

Multiple imputation method (Rubin (1987)), essentially, involves approximating the above posterior density, by drawing independent values, $d^{*(l)}_{mis}, l = 1, 2 \ldots, M$, from the predictive distribution, $Pr(D_{mis}|d_{obs}, E)$, and then constructing,

$$p(\theta|d_{obs}, E) \approx \frac{1}{M}\sum_{l=1}^{M} p(\theta|d^{*(l)})$$

where $d^{*(l)} = \{d_{obs}, d^{*(l)}_{mis}\}$ is the l^{th} "completed-data".

Drawing values of θ from this finite mixture approximation is rather straightforward. For instance, draw, say K, values from each of the M completed data posterior distribution of θ and then use the $K \times M$ values of θ to study the features of the posterior distribution. Raghunathan, Berglund and Solenberger (2018) discuss a moment-based approximation to the mixture posterior density using Rubin's framework (1987).

The critical issue, however, is the construction of predictive distribution for the missing values. A straightforward approach is to develop a joint distribution of all the variables and then construct the conditional predictive distribution of the missing set of values given the observed set of values. However, hundreds of different types of variables may be involved in these data sets. Furthermore, there may be structural dependencies across of variables (such as years smoked and age, repeated measures of heights of children etc, as well as bounds for the missing values) are common occurrences. It is nearly impossible to build a joint distribution for these variables in such complex situations. A popular alternative approach is the sequential regression (also called chained equations and flexible conditional specifications)approach (Raghunathan et al (2001)) which specifies p regression models, each using all other variables as predictors (some variable selection and/or dimensionality reduction techniques may be used to fit the regression models). The missing values are then imputed by drawing values from the corresponding predictive distribution. The regression models can be linear or non-linear, generalized linear or non-linear, semi-parametric, nonparametric etc, depending upon the variable with missing values. Imputation are carried out in a cyclical manner to exploit fully the correlation across all the variables. This method has been implemented in many statistical software packages such as SAS, STATA, and R. A particular implementation recommended for a variety of use is IVEware (www.iveware.org).

This paper proposes refinement to this basic methodology by building model checking and evaluation, any refinement of models, if necessary, in an adaptive fashion to ensure that the imputed values exhibit the properties of the observed values, under the missing at random assumption. The underlying goal is to ensure that each completed data set can be viewed as a potential representative sample from the population. Any external information available will be incorporated in the model building process. All these efforts are integrated using machine learning framework. Validation of the model is also built in the sequential process using the frequentist approach.

The fundamental goal of the proposed procedure is to ensure that

$$Pr(V_{i,mis}|D_{obs,-i}, E) \approx Pr(V_{i,obs}|D_{obs,-i}, E)$$

where V_i is the variable being imputed, $V_{i,mis}$ are the imputed values, $V_{i,obs}$ are the observed values and $D_{obs,-i}$ is the collection of all the observed values in all variables, other than V_i. This should be satisfied for all the variables in the appended data sets at every cycle of sequential regression modeling efforts. The adaptive nature of this proposed procedure is rather obvious. An example of combining data from two surveys and administrative data sources is used to illustrate the methodology. For more details see Bondarenko and Raghunathan (2016). Multiple measures of the distances between the above two distributions are used. A simulation study evaluates the repeated sampling properties of estimates derived from the imputed data sets.

References

1. Bondarenko, I. and Raghunathan, T. E. (2016). Graphical and mumerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in Medicine*, 35, 3007-3020.

2. Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). Bayesian data analysis. CRC Press, Boca Raton, Florida, USA.

3. Raghunathan, T. E., Lepkowski, J. L., Van Hoewyk, J. H., Solenberger, P. W. (2001). A multivariate technique for imputing the missing values using a sequence regression models, *Survey Methodology*, 27, 85-95.

4. Raghunathan, T. E., Berglund, P., and Solenberger, P. (2018). Multiple Imputation in Practice: With Examples using *IVEware*. CRC Press. Boca Raton, Florida, USA.

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

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