Imputation of continuous variables missing at random using the method of simulated scores

Calzolari, Giorgio and Neri, Laura

Universita' di Firenze, Italy., Universita' di Siena, Italy

2002

Imputation of Continuous Variables Missing at Random using the Method of Simulated Scores[1]

Giorgio Calzolari [2] and Laura Neri [3]

[2] Dipartimento di Statistica "G.Parenti", Università di Firenze
Viale Morgagni 59, I-50134 Firenze, Italy
[3] Dipartimento di Metodi Quantitativi, Università di Siena
Piazza San Francesco 8, I-53100 Siena, Italy

Abstract. For multivariate datasets with missing values, we present a procedure of statistical inference and state its "optimal" properties. Two main assumptions are needed: (1) data are missing at random (MAR); (2) the data generating process is a multivariate normal linear regression. Disentangling the problem of convergence of the iterative estimation/imputation procedure, we show that the estimator is a "method of simulated scores" (a particular case of McFadden's "method of simulated moments"); thus the estimator is equivalent to maximum likelihood if the number of replications is conveniently large, and the whole procedure can be considered an optimal parametric technique for imputation of missing data.

Keywords. Simulated scores, missing data, estimation/imputation, structural form, reduced form

1 Introduction

Empirical research in economic and social science often suffer from missing data.

There are three major problems created by missing data. First, if the nonrespondents are systematically different from the respondents and we do not take into account the difference, analysis may be biased. Second, missing data imply loss of information, so estimates will be less efficient than planned. Third, besides theoretical problems like bias and efficiency, a big practical problem is that tools for effectively treating datasets affected by missing values are not readily available. Several methods have been (and continue to be) developed to draw inference from data sets with missing data (Little and Rubin, 1987). The most attractive for applied researcher are those forcing the incomplete dataset into a rectangular complete-data format, where missing values are filled by imputation of suitable estimates.

The main applied approach to the general problem of obtaining valid inferences when facing missing data is the *Multiple Imputation* technique. This idea was explicitly proposed in Rubin (1978) and a decade later the basic reference textbook was published (Rubin, 1987). The currently available solution to this problem is to create multiple imputations specifying one "encompassing multivariate model" for the entire data set (at least conditional on completely observed variables), and then using fully principled likelihood/bayesian techniques for analysis under that

[1] We are grateful to Monica Billio, Fabio Corradi, Emanuela Dreassi, Giampiero M. Gallo, James Lepkowsky, Fabrizia Mealli, Donald B. Rubin and Federico M. Stefanini for suggestions and comments, but retain full responsibility for the contents of this paper.
We gratefully acknowledge financial support from CNR and MURST-MIUR through projects "Stochastic models and simulation methods for dependent data" and "Employment and unemployment in Italy: measurement issues and behavioural analyses".

	X	Y_1	Y_2	...	Y_p
i=1					?
2				?	
.					
.					?
.		?	?	?	?
.					
n			?	?	

Table 1. Dataset with missing values.

model. This generates a posterior distribution for the parameters of the model and a posterior predictive distribution for the missing values (given the model specifications and the observed data). The primary example of such approach is Schafer's freeware, based on Schafer (1997), which involves iterative Markov Chain Monte Carlo (MCMC) computations; Rubin (2000) explains the advantage and disadvantage of such a method. Among the disadvantages, we point out on the fact that iterative versions of software for creating multiple imputations are not always immediately usable for real applications by the typical analyst dealing with missing data; it often needs experts to face with potentially misleading "non convergent" MCMC and other possible difficulties.

In this paper we introduce a method, feasible for data analysts, for creating imputations when dealing with a general missing data *pattern* of continuous variables assuming that missing data are MAR (Missing At Random, Rubin, 1976), and the data generating process is a multivariate normal linear regression. We can obtain the multiple imputed datasets by repeating several times (each time till convergence) the iterative "least-squares estimation/multivariate normal imputation" procedure. We show that, at convergence of the iterative procedure, the estimator is a "method of simulated scores" (MSS, see Hajivassiliou and McFadden, 1990, a particular case of McFadden's (1989) method of simulated moments MSM).

For precision's sake, we must say that through the paper we always refer to a "single" imputation; this makes explanations easier, without loss of generality.

Analytical details on the iterative estimation/imputation process, as well as detailed proofs of the propositions, can be found in Calzolari and Neri (2002).

2 Estimation/imputation based on structural form

A schematic representation of an incomplete dataset is shown in Table 1 where the n rows represent the observational units, X and Y represent variables recorded for those units; X denotes the column(s) of complete variables; in the p columns of Y, question marks identify missing values (they can occur anywhere, in any *pattern*). Our hypothesis through this paper is that the missing data are MAR.

Formally, Y denotes an $n \times p$ matrix of data, Y_{obs} denotes the observed portion of Y, Y_{mis} denotes the missing portion (so that $Y = [Y_{obs}, Y_{mis}]$), X is an $n \times k$ matrix without missing values. When considering a particular column of Y, the portion (rows) of X, corresponding to its missing values, is indicated as X_{mis}.

The task of generating imputations is often a hard task, except in some simple cases such as datasets with only one variable affected by missing values or very special patterns of missingness. The main difficulty is to find a solution for imputing a general pattern of missing data preserving the original association structure of the data. In this paper we deal with this problem when the Y matrix is composed of variables defined on a continuous scale, and when it may be reasonable to use the multivariate normal model to create imputations (Rubin, 1987).

We consider a set of normal linear regression models. We solve the technical problem introducing some convenient modifications into the "sequential regres-

sion multivariate imputation" (SRMI) method by Raghunathan, Lepkowsky, Van Hoewyk, and Solenberger (1997), which is adopted by the imputation software (*IVE-ware*). They build the imputed values by fitting a sequence of regression models and drawing values from the corresponding predictive distribution, under the hypothesis of MAR mechanism, infinite sample size and simple random sampling. The method follows a bayesian paradigm. Each imputation consists of c "rounds". Round 1 starts regressing the variable with the fewest number of missing values, say Y_1, on X, and imputing the missing values with the appropriate regression model. Assuming a flat prior for the regression coefficient, the imputations for the missing values in Y_1 are drawn from the corresponding posterior predictive distribution. After Y_1 has been completed, the next variable with the fewest number of missing values is considered, say Y_2; observed Y_2 values are regressed on (X, Y_1) and the missing values are imputed, and so on. The imputation process is then repeated in rounds 2 through c, modifying the predictor set to include all the Y variables except the one used as the dependent variable. Repeated cycles continue for a pre-specified number of rounds, or until stable imputed values occur (convergence in distribution).

The method we propose follows the SRMI method, but introduces a convenient modification of the variance covariance matrix estimator. Practically, the procedure starts exactly as the SRMI (round 1): we estimate the coefficients of the linear regression model related to the variable with fewest missing values (let be Y_1), by OLS, using the Y_1 observed part ($Y_{obs,1}$). Supposing that $\widehat{\Pi}_1$ is the estimated regression coefficient and $\widehat{\sigma}_{11}$ the residual variance, then the imputed value set is

$$\widetilde{Y_1} = X_{mis,1}\widehat{\Pi}_1 + \sqrt{\widehat{\sigma}_{11}}\ \widetilde{u}_1$$

where $\widetilde{u}_1$ is a vector of independent pseudo-random standard normal deviates.

So we have a first set of completed values for Y_1 and we attach it as an additional column to X. We then regress the next variable with fewest missing values (say $Y_{obs,2}$) against X and the completed Y_1 and use the OLS estimated coefficients and variance for an imputation step that completes Y_2. Going on, the first round ends when all the missing values are completed. As the SRMI's authors put in evidence, the updating of the right hand side variables after imputing the missing values depends on the order in which we select the variables for imputation. Thus, the imputed values for Y_j involve only (X, Y_1, ..., Y_{j-1}), but not Y_{j+1}... Y_p. For this reason the procedure continues to overwrite the imputations for the missing values iteratively. In any iteration after the first round, we always have complete data for all variables, part of which are observed, part have been imputed in the previous iteration. The system of regression equations has, as dependent variable for each equation, the variable to be "imputed if missing" and has on the right hand side all the others variables

$$\begin{aligned}
Y_1 &= X\gamma_{11} + Y_2\gamma_{12} + Y_3\gamma_{13} + ... + Y_p\gamma_{1p} + \varepsilon_1 \\
Y_2 &= X\gamma_{21} + Y_1\gamma_{22} + Y_3\gamma_{23} + ... + Y_p\gamma_{2p} + \varepsilon_2 \\
&... \\
Y_p &= X\gamma_{p1} + Y_1\gamma_{p2} + Y_2\gamma_{p3} + ... + Y_{p-1}\gamma_{pp} + \varepsilon_p
\end{aligned} \tag{1}$$

The $\gamma_{11}, \gamma_{21}, ..., \gamma_{p1}$ are scalars or $(k \times 1)$ vectors depending on X being a single column or a $(n \times k)$ matrix, while all the other γ-s are scalars and the ε-s have a cross-equations multivariate normal distribution.

Equations (1) represent a system of simultaneous equations in structural form. The jointly dependent variables Y appear also on the right hand side of the equations, while the variables X play the role of "exogenous" variables. Such a system is obviously underidentified, as it violates the order condition for identification

(eg. Greene, 2000, sec. 16.3.1): infinite sets of γ-values would be observationally equivalent. It is therefore useless (or impossible) to apply estimation techniques suitable for simultaneous equation systems, like two or three stage least squares, full information maximum likelihood, etc. Nevertheless we can estimate each equation separatly by OLS as in SRMI approach. After coefficients have been estimated by OLS, we compute from residuals the estimate of the $(p \times p)$ variance covariance matrix, say $\widehat{\Psi}$. Differently from the SRMI method, we use the Cholesky decomposition of the matrix $\widehat{\Psi}$ to produce vectors of pseudo-random numbers for imputation, thus considering also covariances besides variances.

When a value of Y_1 is missing, we impute the value obtained from the right hand side of the first equation in (1) where: the γ -s are at the previous iteration estimated value; the value(s) of X is (are) observed; the values of the Y on the right hand side are in any case complete (some of them are observed, the others have been imputed in the previous iteration); the value of ε_1 is "jointly" produced with ε_2, ..., with ε_p by the pseudo-random generator with a cross-equations variance-covariance matrix equal to the last estimated $\widehat{\Psi}$. The same is done for the second equation in (1), filling missing values of Y_2, and so on.

Repeated cycles continue until convergence on the estimated parameters has been achieved.

A question naturally arises: why and when does the the iterative estimation/imputation procedure converge? (Even if the MCMC context is different from ours, still recently Horton and Lipsitz, 2001, p. 246 point out that convergence "remains more of an art form than a science"). Considering the procedure as it has just been described, the answer does not result obvious. In order to answer this question, it is convenient to think that the structural form system can be easily transformed in a reduced form (Greene, 2000, sec. 16.2.3) and to think at the sequence of iterations in a different order, as if iterations were "grouped". Let's first see what happens if we keep parameter values fixed (the γ -s and the Cholesky decomposition of the matrix $\widehat{\Psi}$), and we only iterate substitutions of imputed values on the right hand side of equations (1). These iterated substitutions (e.g. Thisted, sec. 3.11.2) are "exactly" the steps of the well known Gauss-Seidel method for the "simultaneous solution" of the system of equations (also called by econometricians "stochastic simulation", because of the presence of the ε terms). The simultaneous solution, using econometric terminology, is the well known derivation of the "reduced form" (or "restricted" reduced form) from the "structural form". The reduced form system has variables Y only on the left hand side, while the right hand side includes only the variable(s) X and the error terms. Thus, till we hold parametes fixed at some values, the iterated substitution of imputed values will converge to the reduced form derived from the structural form (or restricted reduced form). Now we can re-estimate parameters (with OLS on the structural form) and start again a new cycle of iterated substitutions in (1), and so on.

The strictly thighted sequence of estimations and imputations, for each structural equation, of the SRMI method has thus been reordered, disentangled and converted into a sequence of iterations that are conceptually much more manageble. In each iteration, an OLS estimation of "all" the structural form equations (1), using observed and previously imputed values, is followed by the "simultaneous solution", i.e. transformation into reduced form, that produces "all" the values of the variables Y which are then imputed. Studying the convergence of this new sequence of estimation and imputation phases becomes more manageable, as it will be clear in Section 3.

The SRMI method and the one just proposed follow different paradigms. The former is based on the bayesian paradigm and the latter on the frequentist paradigm. Beyond this difference, it is important to put in evidence the main

technical difference. The SRMI method draws the random normal deviates of the imputation step for each equation "independently"; the method we propose considers stochastic terms drawn from a multivariate normal distribution with a cross-equations variance-covariance matrix ($\widehat{\Psi}$) estimated from residuals.

3 Properties of the estimator

The "good" properties of the estimator discussed above are ensured by the following propositions:

Proposition 1 *For a complete data set, the reduced form parameters estimator, derived from the the* OLS *estimator of the strucural form parameters (γ-s and Ψ), is equal to the* OLS *direct estimator of the reduced form parameters.*

Proposition 2 *The* OLS *estimator of the reduced form parameters, at convergence of the estimation/imputation procedure, is a* MSS *(Method of Simulated Scores) estimator with "one" replication (Hajivassiliou and McFadden, 1990).*

We briefly discuss the implications and consequences of the two propositions; proofs and a more detailed discussion can be found in Calzolari and Neri (2002).

Working on the structural form system (1) has the advantage of being computationally simple as well as rather intuitive. The discussion on convergence of the iterated imputations with fixed parameters (section 2) and *Proposition 1* ensure that we can get exactly the same results if we work directly on the reduced form, estimating its parameters directly by OLS, and using such an estimated reduced form for imputation. However, even if the estimation phase would be simple (even simpler than for (1)), the imputation phase would be much more complex. For each pattern of missing data we should, in fact, specify the appropriate imputation function, with pseudo-random errors that should be conditional on the Y-s observed in that pattern. Since there are 2^p possibly different patterns of missingness, the technical solution would be very hard. Also, there would be no substantial simplifications over the exact maximum likelihood approach, where up to 2^p different conditional densities should be specified, according to which Y-s are observed in each pattern. That's why it is preferable to work, in practice, with the structural form (1).

Nevertheless, passing to the reduced form is necessary for our proof, because the reduced form is much more manageable from the analytical point of view. The reduced form is a simple system of multivariate normal linear regression equations, without endogenous variables on the right hand side. With a complete data set, the OLS normal equations are exactly the same as: *score=0* (this holds both for coefficients and covariance matrix of the reduced form). But since data have been completed by imputation (simulation), the score is, in fact, a "simulated score". As a consequence, we have *Proposition 2*. "One" replication means that only "one" set of pseudo-random error terms are generated and used for imputation. As a consequence, the (asymptotic) variance of the estimated parameters is larger than for maximum likelihood; however, if we perform the same procedure with more replications, the variance decreases, and the simulated scores estimator can reach the maximum likelihood efficiency if the number of replications is conveniently large (in principle, infinitely large).

4 Conclusion

In this paper we have introduced a method for imputation of missing data, assuming data generated by a multivariate normal linear regression model and a MAR missing data mechanism. The method is based on an iterated estimation/imputation procedure. Besides its technical simplicity and feasibility, the peculiarity of the method is in the properties of the estimator. First of all the parameters estimator is consistent and asymptotically normal. Moreover, being

a simulated scores estimator, its efficiency can be improved by increasing the number of replications. Finally, the estimator becomes as efficient as maximum likelihood if the iterative procedure is replicated a sufficiently large number of times, each time iterating to convergence.

We dealt with the missing data problem in the context of a linear normal model in which some observations of some variables (treated as "endogenous" variables) were missing, while other variables (treated as "exogenous" variables) were completely observed.

The imputation approach described can be used to create a single imputation with a variance estimation procedure taking into account the uncertainty due to missing data or can be part of a framework of multiple imputation.

Of course, in many real cases missing data do not affect only continuous variables. The problem exists also for categorical data, count data, or censored variables: generalization of the method here proposed is left to future research.

References

Calzolari G., Neri L. (2002): "A Method of Simulated Scores for Imputation of Continuous Variables Missing At Random", *Quaderni del Dipartimento di Statistica "G. Parenti"* No. 49, Università degli Studi di Firenze.

Gourieroux C., Monfort A. (1996): *Simulation-Based Econometric Methods*. Oxford University Press.

Greene W. H. (2000): *Econometric Analysis* (fourth edition). Upper Saddle River, NJ: Prentice-Hall, Inc.

Hajivassiliou V., McFadden D. (1990): "The Method of Simulated Scores, with Application to Models of External Debt Crises", Cowles Foundation Discussion Paper No. 967, Yale University.

Horton N. J. and Lipsitz S. R. (2001): "Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables", *The American Statistician*, 55, 244-255.

Little R. J. A., Rubin D. B. (1987): *Statistical Analysis with Missing Data*. New York: Wiley.

Mc Fadden D. (1989): "A Method of Simulated Moments for Estimation of Discrete Response Model without Numerical Integration", *Econometrica*, 57, 995-1026.

Raghunathan T. E.: www.isr.umich.edu/src/smp/ive.

Raghunathan T. E., Lepkowski J., Van Voewyk J., Solenberger P. (1997): "A Multivariate Technique for Imputing Missing Values Using a Sequence of Regression Models", Technical Report, Survey Methodology Program, Survey Research Center, ISR, University of Michigan.

Rubin D. B. (1976): "Inference with Missing Data", *Biometrika*, 63, 581-592.

Rubin D. B. (1978): "Multiple Imputations in Sample Surveys-A Phenomeno-Logical Bayesian Approach to Nonresponse", *The Proceeding of the Survey Research Methods Section of the American Statistical Association*, 20-34, with discussion and reply.

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

Rubin D. B. (2000): "The Broad Role of Multiple Imputations in Statistical Science", in *Proceeding in Computational Statistics, 14th Symposium, Utrecht-The Netherlands, 2000*, ed. by J. G. Bethlehem and P. G. M. van der Heijden. Vienna: Physica-Verlag, 3-14.

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

Thisted, R. A. (1988): *Elements of Statistical Computing*. New York: Chapman and Hall.