Two algorithms for producing multiple imputations for missing data are evaluated with simulated data. Software using a propensity score classifier with the approximate Bayesian bootstrap produces badly biased estimates of regression coefficients when data on predictor variables are missing at random or missing completely at random. On the other hand, a regression-based method employing the data augmentation algorithm produces estimates with little or no bias.

Multiple Imputation for Missing Data

A Cautionary Tale

PAUL D. ALLISON
University of Pennsylvania

Multiple imputation appears to be one of the most attractive methods for general-purpose handling of missing data in multivariate analysis. The basic idea, first proposed by Rubin (1977) and elaborated in his 1987 book, is quite simple:

1. Impute missing values using an appropriate model that incorporates random variation.
2. Do this M times (usually three to five times), producing M "complete" data sets.
3. Perform the desired analysis on each data set using standard complete-data methods.
4. Average the values of the parameter estimates across the M samples to produce a single-point estimate.
5. Calculate the standard errors by (a) averaging the squared standard errors of the M estimates, (b) calculating the variance of the M parameter estimates across samples, and (c) combining the two quantities using a simple formula (given below).

Additional discussions of multiple imputation can be found in Little and Rubin (1989) and Landerman, Land, and Pieper (1997).

Multiple imputation has several desirable features:

- Introducing appropriate random error into the imputation process makes it possible to get approximately unbiased estimates of all para-

SOCIOLOGICAL METHODS & RESEARCH, Vol. 28 No. 3, February 2000 301-309

meters. No deterministic imputation method can do this in general settings.
- Repeated imputation allows one to get good estimates of the standard errors. Single imputation methods do not allow for the additional error introduced by imputation (without specialized software of very limited generality).
- Multiple imputation can be used with any kind of data and any kind of analysis without specialized software.

Of course, certain requirements must be met for multiple imputation to have these desirable properties. First, the data must be missing at random, meaning that the probability of missing data on a particular variable Y can depend on other observed variables, but not on Y itself (controlling for the other observed variables). Second, the model used to generate the imputed values must be "correct" in some sense. Third, the model used for the analysis must match up, in some sense, with the model used in the imputation. All these conditions have been rigorously described by Rubin (1987, 1996).

The problem is that it is easy to violate these conditions in practice. There are often strong reasons to suspect that the data are not missing at random. Unfortunately, not much can be done about this. Although it is possible to formulate and estimate models for data that are not missing at random, such models are complex, untestable, and require specialized software. Hence, any general-purpose method will necessarily invoke the missing at random assumption.

Even when the missing at random condition is satisfied, producing random imputations that yield unbiased estimates of the desired parameters is not always easy or straightforward. In this article, I examine two approaches to multiple imputation that have been incorporated into widely available software. We will see that one of them (embodied in software currently retailing for $895) does a terrible job at producing imputations for missing data on predictor variables in multiple regression analysis. The other (available free on the web) does an excellent job.

To compare these two algorithms, I generated 10,000 observations on three variables: X, Y, and Z. Because the primary concern is with bias, the large sample size is designed to minimize sampling variations. X and Z were drawn from a bivariate standard normal distribution with a correlation of .50. Thus, in the observed data, X and Z each

TABLE 1: Estimates for Coefficients in Regression of *Y* on *X* and *Z*

Missing Data Mechanism		*1 Ordinary Least Squares on Original Data*	*2 Listwise Deletion*	*3 Multiple Imputation With SOLAS*	*5 Multiple Imputation With Data Augmentation*
Missing completely at random	*X*	0.979 (.011)	0.969 (.016)	1.141 (.016)	0.976 (.012)
	Z	1.014 (.012)	1.029 (.017)	0.667 (.020)	1.028 (.016)
Missing at random (dependent on *X*)	*X*	1.012 (.012)	0.986 (.025)	1.470 (.013)	1.005 (.025)
	Z	1.007 (.012)	1.011 (.017)	0.448 (.015)	0.997 (.016)
Missing at random (dependent on *Y*)	X	0.993 (.012)	0.695 (.015)	1.350 (.013)	0.985 (.021)
	Z	1.001 (.012)	0.708 (.015)	0.746 (.023)	0.997 (.013)
Nonignorable (dependent on *Z*)	X	1.003 (.012)	0.995 (.016)	1.250 (.013)	1.154 (.015)
	Z	1.002 (.012)	1.007 (.024)	1.215 (.027)	1.245 (.020)

NOTE: Standard errors are in parentheses.

have means of about 0, standard deviations of about 1.0, and a correlation of about .5. *Y* was then generated from the equation

$$Y = X + Z + U,$$

where *U* was drawn from a standard normal distribution, independent of *X* and *Z*.

I then caused about half of the *Z* values to be "missing" according to four different mechanisms:

1. Missing completely at random: *Z* missing with probability .5, independent of *Y* and *X*.
2. Missing at random, dependent on *X*: *Z* missing if $X < 0$.
3. Missing at random, dependent on *Y*: *Z* missing if $Y < 0$.
4. Nonignorable: *Z* missing if $Z < 0$.

For each missing data mechanism, new draws were made from the underlying distribution of *X*, *Z* and *Y*. Column 1 of Table 1 shows results from ordinary least squares regression using the original data with no missing cases on *Z*.

Column 2 of Table 1 shows the results of an ordinary least squares regression using listwise deletion (complete case analysis) on the data set with missing values. Listwise deletion shows little or no bias for all missing data mechanisms except where missingness on *Z* is dependent on *Y*. In that case, both regression coefficients are about 30 percent

lower than the true values. Not surprisingly, the standard errors are always somewhat larger than those based on the full data set. Remarkably, listwise deletion does well even when missingness on Z depends on Z itself (the nonignorable case). More generally, it can be shown that listwise deletion produces unbiased regression estimates whenever the missing data mechanism depends only on the predictor variables, not on the response variable (Glynn 1985; Little 1992).

Next, we will use multiple imputation as implemented by the Windows program SOLAS (Statistical Solutions, http://www.Statsolusa.com). SOLAS (version 1.1) incorporates a number of imputation methods, the most sophisticated of which is multiple imputation using what is described in promotional material as "the latest industry-accepted techniques." In fact, SOLAS uses the algorithm proposed by Lavori, Dawson, and Shera (1995). For our example, the algorithm amounts to this:

1. Do a logistic regression in which the dependent variable is whether or not Z is missing. Independent variables can be chosen by the analyst; I chose both Y and X, although the results are not much different if only one of the two is chosen.
2. Use the estimated logistic model to calculate a predicted probability of being missing. This is called the propensity score, following the work of Rosenbaum and Rubin (1983).
3. Sort the observations by propensity scores and group them into five groups (quintiles).
4. Within each quintile, there are r cases with Z observed and m cases with Z missing. From among the r fully observed cases, draw a simple random sample of r cases with replacement. For each missing case, randomly draw one value (with replacement) from the random sample of r cases and use the observed value of Z as the imputed value. This method is called the approximate Bayesian bootstrap (Rubin and Schenker 1986).

Steps 1 through 4 produce a single imputed data set. Step 4 is repeated M times to produce M complete data sets. For this example, I chose $M = 5$.

I then performed ordinary least squares multiple regressions of Y on X and Z in each of the five completed data sets. The coefficients shown in column 3 of Table 1 are the means of the five estimates. The standard error is obtained from Rubin's (1987) formula. Let b_k be the

estimated regression coefficient in sample k of the M samples, and let s_k be its estimated standard error. The mean of the b_k is $\bar{b}$; its estimated standard error is given by

$$\sqrt{\frac{1}{M}\sum_k S_k^2 + \left(1+\frac{1}{M}\right)\left(\frac{1}{M-1}\right)\sum_k \left(b_k - \bar{b}\right)^2}.$$

In words, this is the square root of the average of the sampling variances plus the variance of coefficient estimates (multiplied by a correction factor $1 + 1/M$).

Results for SOLAS are shown in column 3 of Table 1. All coefficient estimates are clearly biased, and in most cases the biases are quite severe. The worst case is when missingness on Z depends on X, where both coefficients are about 50 percent off, either too high or too low.

Next, we turn to a quite different method for generating multiple imputations, described by Schafer (1997) and embodied in NORM, Schafer's freeware Windows program (http://www.stat.psu.edu/~jls/). To understand the method, consider the following simplified algorithm for random regression imputation:

1. Regress Z on X and Y using all cases with data present on Z. Based on this regression, let $\hat{Z}$ represent the predicted values for all cases including those with data missing and let $\hat{\sigma}$ be the root mean squared error.
2. For cases with data missing on Z, compute $Z^* = \hat{Z} + \hat{\sigma}U$, where U represents a random draw from a standard normal distribution. For cases with no missing data, $Z^* = Z$.
3. Regress Y on X and Z^* for all cases.

Steps 2 and 3 may then be repeated M times. Although this algorithm does a pretty good job, it is not "proper" in Rubin's (1987) sense because it does not adjust for the fact that $\hat{\sigma}$ and the coefficients used to produce $\hat{Z}$ are only estimates, not the true values. This is also true of the multiple imputation methods available in the recently released missing data module for SPSS and the predictive mean matching method described by Landerman et al. (1997). NORM goes the extra mile by using a Bayesian method known as data augmentation to iterate between random imputations under a specified set of parameter values and random draws from the posterior distribution of the para-

meters (given the observed and imputed data). The algorithm is based on the assumptions that the data come from a multivariate normal distribution and are missing at random. A similar algorithm that uses a different method for generating random draws of the parameters is described by King et al. (1999) and incorporated into their freeware program AMELIA (http://gking.harvard.edu/stats.shtml).

Rather than using NORM, I have coded Schafer's (1997) algorithms into two macros for the SAS system, MISS and COMBINE (http://www.ssc.upenn.edu/~allison). The MISS macro was used to produce five complete data sets. After performing regression runs in each data set, the COMBINE macro produced the final estimates shown in column 5 of Table 1. All the estimates are close to the true values except, as expected, for the nonignorable case, in which both coefficients show some upward bias.

Somewhat surprisingly, the estimated standard errors for multiple imputation in column 5 are only slightly smaller than those produced by listwise deletion in column 2, which discarded half the cases. To further investigate the relative performance of listwise deletion and multiple imputation (using the data augmentation algorithm), I simulated 500 random samples, each of size 500, from the model described earlier. In each sample, I made values of Z missing by mechanism 2—missing at random whenever $X < 0$, a condition under which both listwise deletion and multiple imputation are at least approximately unbiased. Both methods were essentially unbiased across the repeated samples. However, the sampling variance of the multiple imputation estimates was considerably smaller: 39 percent less for the coefficient of X and 28 percent less for the coefficient of Z. Within-sample estimation of the standard errors for multiple imputation was also good. The standard deviations across the 500 samples for the two coefficients were .0853 for X and .0600 for Z. By comparison, the average of the within-sample standard error estimates were .0835 for X and .0664 for Z. Nominal 95-percent confidence intervals for the multiple imputation estimates were computed by taking the point estimate plus or minus 1.96 times the estimated standard error. For the X coefficient, 96.4 percent of the confidence intervals included the true value. For Z, the coverage rate was 96.6 percent.

DISCUSSION

We have seen that multiple imputation using the data augmentation algorithm produces reasonably good results, but those produced by SOLAS are severely biased, even when the data are missing completely at random. Why? There is nothing intrinsically wrong with the approximate Bayesian bootstrap algorithm that SOLAS uses to generate the missing values. The problem stems from the fact that when forming groups of "similar" cases, SOLAS only uses information from covariates that are associated with whether or not the data are missing. It does not use information from the associations among the variables themselves. For example, under the missing completely at random condition, missingness on Z is not related to either X or Y. Consequently, the constructed quintiles of "similar" cases are essentially random groupings of the observations. One might as well have imputed Z with random draws from all the observed values of Z. By contrast, the data augmentation method does use information on both X and Y to generate the imputed values. In short, SOLAS violates Rubin's principle that one must use a "correct" model in producing the imputed values.

The failure of SOLAS under the missing at random conditions points out another crucial requirement for correct imputation. Under condition 2, when missingness on Z is highly associated with X, the imputed values of Z generated by SOLAS do make use of the correlation between X and Z, but Y contributes nothing to the imputations. To get unbiased estimates in a regression analysis, it is essential to use the dependent variable to impute values for missing data on the predictor variables (Schafer 1997). Failure to include the dependent variable implies that the imputed values for the predictors will not be associated with the dependent variable, net of other variables in the imputation model. Consequently, regression coefficients for predictor variables with large fractions of imputed data will show substantial biases toward zero (Landerman et al. 1997). Although it might seem that using the dependent variable to impute predictors would produce inflated coefficient estimates, this is avoided by the addition of the random component to the imputed value.

Are there any conditions under which SOLAS lives up to its claims? Its imputation algorithm, as originally described by Lavori et al.

(1995), was designed for a randomized experiment with repeated measures on the response variable. However, some participants dropped out of the study before all response measurements could be made. The goal was to impute the missing responses based on previous response measurements, as well as baseline covariates. While I have not done any simulations for this kind of design, the results reported by Lavori et al. indicate that their method performs well in this situation.

What is clear is that the method is not helpful in imputing predictor variables. Indeed, listwise deletion is clearly better in all the conditions studied here. Executives at Statistical Solutions Limited, the originator of SOLAS, have been made aware of this problem and state that a new version with improved algorithms is currently in preparation (Aidan McDonnell, personal communication, October 18, 1999). Nevertheless, the current version (1.1) of the software is still heavily marketed, with no warning about the dangers of imputing missing predictors.

REFERENCES

Glynn, Robert. 1985. "Regression Estimates When Nonresponse Depends on the Outcome Variable." D.Sc. dissertation, Harvard University School of Public Health, Cambridge, MA.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 1999. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation." Unpublished manuscript. Available at http://gking.harvard.edu/stats.shtml

Landerman, Lawrence R., Kenneth C. Land, and Carl F. Pieper. 1997. "An Empirical Evaluation of the Predictive Mean Matching Method for Imputing Missing Values." *Sociological Methods & Research* 26:3-33.

Lavori, P., R. Dawson, and D. Shera. 1995. "A Multiple Imputation Strategy for Clinical Trials With Truncation of Patient Data." *Statistics in Medicine* 14:1913-25.

Little, Roderick J. A. 1992. "Regression With Missing X's: A Review." *Journal of the American Statistical Association* 87:1227-37.

Little, Roderick J. A., and Donald B. Rubin. 1989. "The Analysis of Social Science Data With Missing Values." *Sociological Methods & Research* 18:292-326.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70:41-55.

Rubin, Donald B. 1977. "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys." *Journal of the American Statistical Association* 72:538-43.

———. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

———. 1996. "Multiple Imputation After 18+ Years [with discussion]." *Journal of the American Statistical Association* 91:473-89.

Rubin, Donald B., and Nathaniel Schenker. 1986. "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse." *Journal of the American Statistical Association* 81:366-74.

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

Paul D. Allison is a professor of sociology at the University of Pennsylvania. His recently published books include Multiple Regression: A Primer *and* Logistic Regression Using the SAS System: Theory and Applications. *He is currently writing a book on missing data. Each summer, he teaches 5-day workshops on event history analysis and categorical data analysis.*