Title
Multiple Imputations for LInear Regression Models

Author
Brownstone, David

Publication Date
1991-11-01

Multiple Imputations for Linear Regression Models

David Brownstone

November 1991
Working Paper, No. 37

The University of California
Transportation Center

University of California
Berkeley, CA 94720

The University of California

Transportation Center

The University of California Transportation Center (UCTC) is one of ten regional units mandated by Congress and established in Fall 1988 to support research, education, and training in surface transportation. The UC Center serves federal Region IX and is supported by matching grants from the U.S. Department of Transportation, the California State Department of Transportation (Caltrans), and the University.

Based on the Berkeley Campus, UCTC draws upon existing capabilities and resources of the Institutes of Transportation Studies at Berkeley, Davis, and Irvine; the Institute of Urban and Regional Development at Berkeley; the Graduate School of Architecture and Urban Planning at Los Angeles; and several academic departments at the Berkeley, Davis, Irvine, and Los Angeles campuses. Faculty and students on other University of California campuses may participate in Center activities. Researchers at other universities within the region also have opportunities to collaborate on selected studies. Currently faculty at California State University, Long Beach, and at Arizona State University, Tempe, are active participants.

UCTC's educational and research programs are focused on strategic planning for improving metropolitan accessibility, with emphasis on the special conditions in Region IX. Particular attention is directed to strategies for using transportation as an instrument of economic development, while also accommodating to the region's persistent expansion and while maintaining and enhancing the quality of life there.

The Center distributes reports on its research in working papers, monographs, and in reprints of published articles. For a list of publications in print, write to the address below.

University of California
Transportation Center

108 Naval Architecture Building
Berkeley, California 94720
Tel: 415/643-7378
FAX: 415/643-5456

Authors of papers reporting on UCTC-sponsored research are solely responsible for their content. This research was supported by the U.S. Department of Transportation and the California State Department of Transportation, neither of which assumes liability for its content or use.

Multiple Imputations for Linear Regression Models

David Brownstone

Department of Economics
University of California at Irvine

Working Paper, No. 37

The University of California Transportation Center
University of California at Berkeley

 September 28, 1991

MULTIPLE IMPUTATIONS FOR LINEAR REGRESSION MODELS

David Brownstone*
Department of Economics
U.C. Irvine
Irvine, CA 92717
BITNET: DBROWNST@UCI

Abstract

Rubin (1987) has proposed multiple imputations as a general method for estimation in the presence of missing data. Rubin's results only strictly apply to Bayesian models, but Schenker and Welsh (1988) directly prove the consistency of multiple imputations inferences when there are missing values of the dependent variable in linear regression models. This paper extends and modifies Schenker and Welsh's theorems to give conditions where multiple imputations yield consistent inferences for both ignorable and nonignorable missing data in exogenous variables. One key condition is that the imputed values must have the same conditional first and second moments as the true values. Monte Carlo studies show that the multiple imputation covariance estimates are accurate for realistic sample sizes. They also support the applications of multiple imputations in Brownstone and Valletta (1991), where the multiple imputations estimates substantially changed the qualitative conclusions implied by the model.

* Financial support from the U.C. Irvine Research Unit in Mathematical Behavioral Sciences is gratefully acknowledged. Cheng Hsiao, David Lilien, Ken Small, anonymous referees, and participants at Camp Econometrics III provided many useful comments and suggestions, but they are not reponsible for the remaining flaws.

1. Introduction

Econometricians have been active in developing techniques for handling "nonignorable" missing data such as sample selection and truncation (see Heckman, 1976). There has been far less interest in ignorable missing data (i.e. where the missing data mechanism depends only on observed exogenous data). The two usual methods for dealing with ignorable missing data are to use only cases with complete data or to impute missing values and then treat the imputed values as if they were observed. The former solution is frequently inefficient and the latter solution almost always produces biased confidence intervals and tests.

Little and Rubin (1987) show that two general methods for consistent inferences with ignorable missing data are maximum likelihood and multiple imputations. This study concentrates on the latter since maximum likelihood techniques are more familiar to most econometricians, frequently require strong distributional assumptions, and are also frequently difficult to compute using standard software packages. In contrast, multiple imputation methods are relatively easy to implement. Moreover, some of the imputation methods described in Section 3 of this paper do not require strong distributional assumptions. In principle, the imputations can be done once and then used for many different analyses. Thus, by including the multiple imputations in a public use file, confidential information such as exact addresses could be used to improve the quality of the imputations without sacrificing confidentiality of the data in the public use file.

Rubin (1987, Chapter 4) shows that if the data are being analyzed and missing data being imputed using full Bayesian models, then multiple imputations provide consistent estimates. These results can be difficult to apply in situations where the analyst is not willing (or able) to specify a full Bayesian model. Schenker and Welsh (1988) give a direct proof of the consistency of multiple imputations when

here are missing values of the dependent variable in a linear regression model. This application of multiple imputations is not practically important since the complete data least squares estimator is the maximum likelihood estimator for this model and therefore dominates the multiple imputations estimator. The next section of this paper reviews the multiple imputations method and shows how Schenker and Welsh's results can be modified and extended to provide general conditions for the consistency and asymptotic normality of multiple imputations estimators when there are missing data in independent variables in linear regression models.

When there are missing data in independent variables, or, as in Brownstone and Valletta (1991), additional information which can be used to improve imputations for dependent variables, then multiple imputations will generally be more efficient than the complete data least squares estimators. In these cases the non–missing dependent variable observations corresponding to the observations with missing independent variables provide additional information which is captured by the multiple imputations procedures. However, multiple imputation estimators are generally not fully efficient, as shown in Section 4 by comparing them with Ruud's (1991) Simulated EM estimators. Nevertheless, for the applications in Brownstone and Golob (1992) and Brownstone and Valletta (1991), multiple imputations estimators were substantially more efficient than the corresponding complete data estimators.

The third section considers imputation methods for ignorable missing data in both dependent and independent variables in regression models. Two methods also analyzed by Schenker and Welsh are shown to satisfy the conditions for consistency given in the second section. A new method which uses bootstrap iterations to draw the imputation values is also described and shown to satisfy the consistency conditions. This "bootstrap" imputation method has the advantage of being less sensitive to departures from normality. Monte Carlo studies illustrate the

consistency and small sample performance of multiple imputations using these imputation methods. The results in this section justify the application of multiple imputations in Brownstone and Valletta's (1991) study of measurement errors in cross–section and dynamic earnings equations.

The fourth section discusses the application of multiple imputations to regression models with nonignorable missing data. The key new difficulty presented by this case is obtaining consistent parameter and standard error estimators for the imputation model. This section shows how multiple imputations methods can be used to obtain consistent standard errors for weighted linear regression with estimated weights and a Feasible GLS alternative to Heckman's (1976) two–step estimator for sample selection models. This latter application allows for consistent inference without the complex matrix computations given by Lee, Maddala, and Trost (1980). The finite–sample behavior of the multiple imputations estimates is demonstrated with a Monte Carlo example based on Brownstone and Englund's (1991) model of Swedish housing demand.

2. Multiple Imputation Methods

The fully efficient approach to the problem of missing data is to specify a model for the missing data mechanism and then jointly estimate this model together with the analysis model using maximum likelihood techniques (see Fuller, 1987 and Little and Rubin, 1987). A simpler approach is to somehow generate imputed values for the missing data, and then analyze the resulting completed data set as if there were no missing values. While this is simple, it also leads to downward biased standard error estimates regardless of the accuracy of the imputation procedure. The difficulty with this approach is that some method is needed to account for the errors in the imputation procedure.

Rubin (1987) has proposed multiple imputation as a general method for generating consistent inferences from data sets with imputed values. Instead of just generating one imputation, a number of imputations are created for each missing observation, resulting in a number of completed data sets. Estimators and test statistics are computed from each completed data set and then combined to generate the final inferences. The next section gives explicit methods for obtaining proper multiple imputations. This section summarizes the methods used for combining estimators computed from each completed data set and shows how Schenker and Welsh's (1988) results can be extended to handle missing data in exogenous variables in linear regression models.

Assume that we are interested in estimating some vector θ, and, in the absence of missing data and conditional on all of the observed data, we have an estimator $\hat{\theta}$ which has an asymptotic Normal distribution with mean θ and covariance Ω. Suppose also that there is a consistent estimator, $\hat{\Omega}$, for Ω. Further assume that we have a "proper" imputation model (to be defined later), and that we have drawn a set of M independent (conditional on the observed data) imputations for each missing value. For each of the resulting M completed data sets compute $\hat{\theta}_i^*$ and $\hat{\Omega}_i^*$. The final estimate of θ is the average of the point estimates from the M completed data sets:

$$\bar{\theta}_M = M^{-1} \sum_{i=1}^{M} \hat{\theta}_i^* . \tag{1}$$

If $\bar{\Omega}_M$ is the corresponding average of the completed data covariance estimates and

$$B_M = \sum_{i=1}^{M} (\hat{\theta}_i^* - \bar{\theta}_M)(\hat{\theta}_i^* - \bar{\theta}_M)' / (M-1), \tag{2}$$

then

$$T_M = \bar{\Omega}_M + (1 + M^{-1})B_M \tag{3}$$

is the estimate of the covariance of $(\bar{\theta}_M - \theta)$. Note that T_M can be heuristically derived from:

$$\mathrm{Cov}(\bar{\theta}_M) = \mathrm{E}(\mathrm{Cov}(\bar{\theta}_M | \hat{\alpha})) + \mathrm{Cov}(\mathrm{E}(\bar{\theta}_M | \hat{\alpha})) , \tag{4}$$

where $\hat{\alpha}$ are estimates of the unknown parameters in the imputation model. The first term on the right–hand side of equation (4) is estimated by $\bar{\Omega}_M$, and it represents the covariance within a set of imputations. The second term is estimated by $(1 + M^{-1})B_M$, and it represents the covariance across different sets of imputations.

As both the number of imputations, M, and the sample size get large, the Wald test statistic for the null hypothesis that $\theta=\theta^0$,

$$(\theta^0 - \bar{\theta}_M)' T_M^{-1} (\theta^0 - \bar{\theta}_M)/K , \tag{5}$$

has an asymptotic χ^2_K distribution (K is the rank of θ). If M is finite, but still moderately large ($M \geq 5K$), then Rubin (1987) shows that a better asymptotic approximation to the null distribution of the Wald test is given by an F distribution with K and ν degrees of freedom, where

$$\nu = (M-1)(1+r_M^{-1})^2 \text{ and} \tag{6}$$

$$r_M = (1+M^{-1})\,\mathrm{Tr}(B_M \bar{\Omega}_M^{-1})/K.$$

Note that, if $K = 1$, then r_M is the relative increase in variance due to nonresponse. Li *et. al.* (1991) give an alternative approximation for smaller M. In some

applications, particularly public use files, M must be small. However, it is clear that the variance of B_M is reduced by larger M. This suggests that it is better to compute M large enough so that ν in equation (6) is large enough to use the asymptotic χ^2_K distribution for inference. All of the estimations reported in this paper used M large enough so that ν is greater than 100. The resulting M values are between 10 and 20.

The key issue is how to generate "proper" multiple imputations; i.e. imputation methods where $\bar{\theta}_M$ and T_M are consistent for θ and Ω. Rubin (1987) shows that if one is using an explicit Bayesian model, then making independent draws from the posterior predictive density function for the missing observations will generate proper imputations. Since it can be difficult to verify that a particular imputation procedure is proper without using a formal Bayesian model, I will discuss conditions which are easier to verify for linear regression models.

Consider the standard linear model:

$$Y = X\theta + \epsilon, \tag{7}$$

where, conditional on X, the components of ϵ are independent and identically distributed random variables with mean 0 and variance σ^2, and θ is a K–dimensional vector of unknown coefficients. In the absence of missing data, θ would be estimated by ordinary least squares, $\hat{\theta}$, and inference would be based on:

$$\sqrt{N}\,(\hat{\theta} - \theta) \xrightarrow{d} \Delta^{-1}\,(\lim \frac{X'\epsilon}{\sqrt{N}}) = N(0,\sigma^2\Delta^{-1}) \tag{8}$$
$$s^2 = Y'(I - X(X'X)^{-1}X')Y \,/\, (N-K) \xrightarrow{p} \sigma^2,$$

where $\Delta = \lim (X'X)/N$ is assumed to be positive definite. Now suppose that the

first N_0 observations contain missing data in either (or both) of the exogenous or endogenous variables, but that there are no missing data in the remaining N_1 ($= N - N_0$) observations.

Assume further that there is some method for producing imputed values of the missing variables, denoted Y_0^* and X_0^*, conditional on the observed data and the imputation model parameters, which has the following properties:

$$\text{(9)} \qquad \begin{aligned} \text{Plim } (Y_0^{*\prime} X_0^*)/N &= \text{Plim } (Y_0' X_0)/N \text{ and} \\ \text{Plim } (X_0^{*\prime} X_0^*)/N &= \text{Plim } (X_0' X_0)/N \, . \end{aligned}$$

These conditions, which state that the asymptotic moments of the imputed variables match the first two asymptotic moments of the unobserved true variables, are sufficient to establish the consistency of the multiple imputations parameter estimator, $\bar{\theta}_M$, since the completed data least squares estimate, $\hat{\theta}^*$, is given by:

$$\text{(10)} \qquad \hat{\theta}^* = [(X_0^{*\prime} X_0^*) + (X_1' X_1)]^{-1}[(X_0^{*\prime} Y_0^* + X_1' Y_1)].$$

Assumptions 9 imply that Plim $\hat{\theta}^*$ = Plim $\hat{\theta} = \theta$.

Establishing the asymptotic distribution of $\bar{\theta}_M$ requires additional assumptions about the imputation process. Suppose that:

$$\text{(11)} \qquad \sqrt{N}\,(\hat{\theta}_i^* - \hat{\theta}) \xrightarrow{d} N(0,\Sigma).$$

If the stacked vector $\sqrt{N}\,(\hat{\theta}_\cdot^* - \hat{\theta})$ converges to a multivariate normal distribution with off–diagonal correlations given by

$$\text{(12)} \qquad \text{Plim } N\,(\hat{\theta}_i^* - \hat{\theta})\,(\hat{\theta}_j^* - \hat{\theta})' = \Sigma_c \, ,$$

then for fixed $M \geq 2$,

$$\sqrt{N}\,(\bar{\theta}_M - \hat{\theta}) \xrightarrow{d} N(0,\ (\Sigma/M + \Sigma_c(M-1)/M). \tag{13}$$

Schenker and Welsh's (1988) Lemma 1 then implies that

$$\begin{aligned}\sqrt{N}\,(\bar{\theta}_M - \theta) &= \sqrt{N}\,(\bar{\theta}_M - \hat{\theta}) + \sqrt{N}\,(\hat{\theta} - \theta) \\ &\xrightarrow{d} N(0, \sigma^2\Delta^{-1} + (\Sigma/M + \Sigma_c(M-1)/M))\ .\end{aligned} \tag{14}$$

Since it is clear that $N\bar{\Omega}_M$ is a consistent estimator of $\sigma^2\Delta^{-1}$, the consistency of NT_M for the asymptotic covariance of $\bar{\theta}_M$ depends on $N(1+M^{-1})B_M$. Straightforward calculation shows that

$$\text{Plim } N\,(\hat{\theta}_i^* - \bar{\theta}_M)\,(\hat{\theta}_i^* - \bar{\theta}_M)' = (\Sigma - \Sigma_c)(M-1)/M\ , \text{ so that} \tag{15}$$

$$\text{Plim } N(1+M^{-1})B_M = (1+M^{-1})(\Sigma - \Sigma_c)\ . \tag{16}$$

Comparing equations (14) and (16), it is clear that T_M is consistent only if $\Sigma = 2\Sigma_c$. If this condition is satisfied, then a large number of multiple imputations reduces the variance component estimated by $(1+M^{-1})B_M$ by a factor of 2 relative to a single imputation. The next section will examine some simple imputation methods and demonstrate that they satisfy the $\Sigma = 2\Sigma_c$ condition.

Although the above analysis generalizes Schenker and Welsh (1988), there are some important differences. The key difference is that Schenker and Welsh assume that $\Sigma_c = 0$, and they center their analysis around the least squares estimator for the complete data, $\hat{\theta}_1$. This allows them to get the stronger results that $\sqrt{N}\,(\bar{\theta}_M - \theta)$ is asymptotically independent of B_M and that B_M converges to a Wishart

distribution. Unfortunately, their conditions only apply to the case where there is ignorable missing data in the endogenous variable in a regression model and the missing values are imputed using only the observed data (i.e. Y_1 and X_1). These conditions imply that the complete data estimator, $\hat{\theta}_1$, has lower variance than the multiple imputations estimator, $\bar{\theta}_M$. Schenker and Welsh's results further imply that as M goes to infinity, the asymptotic covariance of $\bar{\theta}_M$ converges to the asymptotic covariance of $\hat{\theta}_1$. Therefore multiple imputations or any other attempt to improve on $\hat{\theta}_1$ is not useful in this situation. The generalizations and modifications carried out in this section justify the use of multiple imputations in situations where $\bar{\theta}_M$ does have lower asymptotic covariance than $\hat{\theta}_1$.

3. Imputation Methods

The previous sections shows that the consistency of the multiple imputations estimators, $\bar{\theta}_M$ and T_M, depends crucially on the properties of the methods used to draw the imputed values. This section describes some simple imputation procedures and shows that they satisfy all of the requirements for the consistency of the multiple imputations estimators given in the previous section. These results will be illustrated by a number of Monte Carlo examples. All of the methods described here assume that the missing data process is ignorable conditional on fully observed exogenous variables. This implies that the complete data least squares estimator, $\hat{\theta}_1$, is consistent. The next section will discuss extensions to cases with nonignorable missing data.

To keep the notation simple, I will first consider the case where there is only one exogenous variable. The general approach to generating imputations which match the first two moments of the missing variable (and therefore satisfy conditions 9) is given by:

$$X_0^* = E(X_0|Y_0) + \eta_0^*, \tag{17}$$

where η_0^* are independent draws from the distribution of $X_0 - E(X_0|Y_0)$. Note that η_0^* have mean 0 and variance equal to $V(X_0|Y_0) = \sigma_\eta^2$, so that

$$E(X_0^{*\prime}Y_0) = E(E(X_0|Y_0)'Y_0) = E(X_0'Y_0) \text{ and} \tag{18}$$

$$\begin{aligned} E(X_0^{*\prime}X_0^*) &= E(E(X_0|Y_0)'E(X_0|Y_0)) + E(E(\eta_0^{*\prime}\eta_0^*)|Y_0) \\ &= V(E(X_0|Y_0)) + E(X_0)'E(X_0) + E(V(X_0'X_0)|Y_0) \\ &= V(X_0'X_0) + E(X_0)'E(X_0) = E(X_0'X_0) \end{aligned} \tag{19}$$

Since the missing data process is ignorable, standard parametric or nonparametric regression methods (see Manski, 1991) can be used to consistently estimate $E(X_1|Y_1)$ and $V(X_1|Y_1)$ from the observed data. These estimates can then be used to create imputed values according to equation (17) above. If, as will be assumed in the rest of this section, (X,Y) are jointly normally distributed, then $E(X_1|Y_1)$ and $V(X_1|Y_1)$ can be estimated by regressing X_1 on Y_1. Note that equations (18) and (19) still hold even if (X,Y) are not joint normal, as long as some consistent estimators of $E(X_1|Y_1)$ and $V(X_1|Y_1)$ are available. It is crucial, however, to condition on Y.

The final step is to establish that the imputations satisfy $\Sigma = 2\Sigma_c$. Without loss of generality we can further assume that $E(X) = E(Y) = 0$. Conditional on the observed data, which includes any estimated parameters in $E(X_0|Y_0)$, $\sqrt{N}(\hat{\theta}_i^* - \hat{\theta})$ has the same asymptotic distribution as

$$(20) \qquad \Delta^{-1}\left[(\eta_i^* - \eta_0)'\,(\epsilon_0 - \theta(E(X_0|Y_0) + \eta_i^*)\right] / \sqrt{N}\,,$$
$$\equiv \Delta^{-1}\,(S_i' R_i)/\sqrt{N}$$

where $\eta_0 = X_0 - E(X_0|Y_0)$. Since we are assuming that (X,Y) are bivariate normal, (S_i, R_j) are also joint normally distributed so that $(S_i' R_j)$ are elements of a matrix with an asymptotic Wishart distribution. Since η_i^* have the same distribution as η_0 and are independent of ϵ_0 and η_0,

$$E(S_i' S_j)/N = \begin{cases} 2\sigma_\eta^2 \text{ if } i=j \\ \sigma_\eta^2 \text{ if } i \neq j \end{cases}$$

$$(21) \qquad E(S_i' R_j)/N = \begin{cases} 0 \quad \text{if } i=j \\ \sigma_{\eta\epsilon}^2 \text{ if } i \neq j \end{cases}$$

$$E(R_i' R_j)/N = \begin{cases} \Lambda + \theta^2\sigma_\eta^2 \text{ if } i=j \\ \Lambda \quad \text{if } i \neq j \end{cases}, \text{ where}$$

$$\Lambda = \sigma_\epsilon^2 + \theta^2 V(E(X_0|Y_0)) - 2\rho_{XY}^2\sigma_\epsilon^2\,.$$

Moment formulas for the Wishart distribution (see Press, 1982, page 115) then give:

$$(22) \qquad E(S_i' R_i)^2/N = 2\Lambda\sigma_\eta^2 + 2\theta^2\sigma_\eta^4$$
$$E((S_i' R_i)\,(S_j' R_j))/N = \Lambda\sigma_\eta^2 + \sigma_{\epsilon\eta}^4\,.$$

Since $\theta = \rho_{XY}\sigma_Y/\sigma_X$, $\sigma_\eta^2 = \sigma_X^2(1-\rho_{XY}^2)$, and $\sigma_{\epsilon\eta}^2 = (\rho_{XY}^3 - \rho_{XY})\sigma_Y\sigma_X$,

$$(23) \qquad E(S_i' R_i)^2/N = 2\,E((S_i' R_i)\,(S_j' R_j))/N\,,$$

and therefore

$$2\Sigma_c = \Delta^{-1}[2\,E((S_i'R_i)(S_j'R_j))/N]\Delta^{-1} \tag{24}$$
$$= \Delta^{-1}[E((S_i'R_i)^2)/N]\Delta^{-1} = \Sigma .$$

Thus, the multiple imputations estimator, $\bar{\theta}_M$, is consistent and asymptotically normally distributed for fixed $M \geq 2$, and T_M is a consistent estimator of its asymptotic covariance.

The computations required for multiple imputations with the imputation procedure in equation (17) are similar to those required by the EM algorithm for maximum likelihood described in Little and Rubin (1987, pp. 143). The E (expectation) step calculates the two "complete data" sufficient statistics conditional on the observed values and current parameter values according to:

$$S_1 = \hat{X}_0'Y_0 + X_1'Y_1 , \quad \hat{X}_0 = E(X_0|X_1, Y, \theta)$$
$$S_2 = \hat{X}_0'\hat{X}_0 + Var(X_0|X_1, Y, \theta) + X_1'X_1 \tag{25}$$

The M (maximization) step calculates a new estimate of θ using the above sufficient statistics. The EM algorithm iterates between the E and M step, using the new θ from the M step to update the sufficient statitics in equations (25).

If the imputed values from equation (17), X_0^*, replace $\hat{X}_0$ and the variance term is dropped, then the resulting simulated sufficient statistics calculated from (25) are clearly unbiased estimates of S_1 and S_2. If this method of updating the sufficient statistics is iterated similarly to the EM algorithm, then it becomes Ruud's (1991) Simulated EM estimator. Of course, the multiple imputation algorithm does not update the parameter estimates before each imputation.

Another difference is that, except for special cases with exponential families (including the slope parameters in the linear model), the completed data estimators used in multiple imputations do not maximize the expected log–likelihood as required by the EM algorithm. Therefore it is clear that the multiple imputation estimator is generally not equal to, nor as efficient as, the maximum likelihood estimator.

The simple model analyzed above is not very interesting from a practical perspective since, as in Schenker and Welsh's model, the multiple imputations estimator is dominated by least squares computed from the complete data, $\hat{\theta}_1$. However, if there are additional fully observed exogenous variables, Z, then, as long as $(X,Y|Z)$ is bivariate normal and $E(X|Y,Z)$ is homoskedastic and linear in Z, the above analysis will show the consistency of multiple imputations if everything is conditioned on Z. If $\hat{\alpha}$ is the least squares estimator of X_1 on Y_1 and Z_1 and s_η^2 is the standard unbiased least squares estimator of the conditional variance, σ_η^2, then one set of proper imputations can be generated from the following procedure: 1) draw σ_η^{2*} from $(N_1-K)s_\eta^2/\chi^2_{(N_1-K)}$ and draw α^* from a $N(\hat{\alpha},\sigma_\eta^{2*}[(Y_1\ Z_1)'(Y_1\ Z_1)]^{-1})$, then 2) construct

$$X_0^* = (Y_0\ Z_0)\alpha^* + F\ \sigma_\eta^*, \tag{26}$$

where F is a vector of N_0 independent draws from a standard normal distribution. Additional sets of imputations needed for multiple imputations can be constructed by repeating the above procedure. Schenker and Welsh call this method, which is a simple extension of a method used in Herzog and Rubin (1983), the "normal imputation" procedure.

The "normal imputation" procedure can be easily modified to accommodate multivariate missing data. If J is the number of variables to be imputed, X_0 is now a $N_0 \times J$ matrix, $\hat{\alpha}$ is a matrix of least squares (or seemingly unrelated regression) estimates with J columns, and s_η^2 is a $J \times J$ estimated residual correlation matrix. $\sigma_\eta^{2^*}$ is drawn from a Wishart$(s_\eta^2, J, (N_1 - K))$ distribution, vec(α^*) is drawn from a $N(\text{vec}(\hat{\alpha}), \sigma_\eta^{2^*} \otimes [(Y_1 \; Z_1)'(Y_1 \; Z_1)]^{-1})$ distribution, and F is a $N_0 \times J$ matrix of independent standard normal random variables. This multivariate imputation procedure clearly also works for imputing missing values of the endogenous variable, Y_0. In this case, the regression(s) used to impute Y_0 only contain Z as right hand (exogenous) variables.

The practical usefulness of the normal imputation procedure is illustrated here with a small Monte Carlo study. The data are generated according to:

$$
\begin{aligned}
ys &= 1 + xs + \varphi_1 \\
x &= xs + \varphi_2 \\
y &= ys - .2xs + \varphi_3 ,
\end{aligned}
\tag{27}
$$

where the φ_i are each composed of 200 independent draws from a standard normal distribution and *xs* is also drawn from a standard normal but held fixed throughout the Monte Carlo repetitions. The last 100 observations of *ys* and *xs* are treated as missing, and they are replace by (multiply) imputed values using the multivariate normal imputation procedure described in the previous paragraphs. This design is a simplified version of a model used in Brownstone and Valletta (1991), where *ys* and *xs* represent true values of primary job earnings and tenure respectively. The true values, obtained from employer administrative records, are only observed in a relatively small validation study, but the reported values, *y* and *x*, are observed in both the validation and main samples.

The Monte Carlo results for the slope coefficient, given in Table 1, are based on 400 Monte Carlo repetitions. As expected, all of the slope estimates are very close to the true value, 1. The multiple imputations variance estimator, T_M, is also quite close to its true value. Table 1 also illustrates the general conclusion that the variance of the multiple imputations estimator lies between the variance of the complete data estimator, $\hat{\theta}_1$, and the estimator computed using the true values of the missing observations, $\hat{\theta}$ (which is not available except in a Monte Carlo study). Although the completed data estimator, $\hat{\theta}^*$ (least squares treating one set of imputed values as fixed) is only slightly less efficient than multiple imputations in this example, the standard error estimates computed using the usual least squares formulas are downward biased by almost 50 percent.

Table 1: Monte Carlo Results For Slope Coefficient and SE Estimators in Regression of *ys* on a constant and *xs*

Estimator	Mean	Standard Deviation
$\bar{\theta}_M$	0.99	.088
$\sqrt{T}_M$	0.084	.008
$\hat{\theta}^*$	0.99	.094
$SE(\hat{\theta}^*)$	0.067	.005
$\hat{\theta}_1$	1.00	.100
$SE(\hat{\theta}_1)$	0.99	.004
$\hat{\theta}$	1.00	.065
$SE(\hat{\theta})$	0.069	.003

Note: $SE(\cdot)$ denotes the standard error of the least squares coefficient estimator using the usual formula $(s^2(X'X)^{-1})$.

When faced with data generated from equations (27), many applied econometricians would use the complete data estimator, $\hat{\theta}_1$, which is consistent, but inefficient. Some would use a single imputation, which, if proper in the sense defined at the beginning of this section, would also yield a consistent estimator. Unfortunately, treating the imputed values as fixed leads to biased inferences. The multiple imputations estimator is relatively easy to compute, more efficient than $\hat{\theta}_1$ and $\hat{\theta}^*$, and yields consistent inferences. Finally, some would treat y and x as proxy variables and estimate the slope coefficient by regressing y on x and a constant. This would be disastrous for the design used here, yielding an average estimate of .41 with a standard deviation of .07.

Additional Monte Carlo experiments were performed using variations on the design in equations (27). As the measurement error (φ_2 and φ_3) variances increase, the variances of the imputation estimators ($\bar{\theta}_M$ and $\hat{\theta}^*$) increase towards the variance of the complete data estimator, $\hat{\theta}_1$. Also, as the number of multiple imputations, M, is reduced to 5 or 10, the variance of the multiple imputations variance estimator, T_M, increases, but its mean value over the Monte Carlo repetitions remains close to the true values. The results of these additional Monte Carlo experiments are reported in a separate appendix available from the author.

Although the above analysis of the normal imputation procedure assumed joint normality of (X,Y), all that is necessary is that the moment conditions in equations (21) and (22) are satisfied. Schenker and Welsh suggest a modification of the normal imputation procedure which is less sensitive to the normality assumption. Their "adjusted normal imputation" method replaces F in equation (26) with N_0 independent draws with replacement from the studentized residuals from the regression of X_1 on (Y_1,Z_1). They then use Freedman's (1981) results on the consistency of bootstrap distributions to show that this adjusted method has the same asymptotic properties as the normal imputation procedure.

One difficulty with the "adjusted normal imputation" method is that it still assumes that $\hat{\alpha}$ and s_η^2 follow a normal and chi–squared distribution, which is only asymptotically correct. This suggests further modifying the imputation procedure to also draw α^* and σ_η^2 from their bootstrap distributions. This "bootstrap imputation" procedure is implemented by:

a) Draw a N_1 element vector of simulated residuals by drawing independently with replacement from the least squares residual vector from the regression of X_1 on (Y_1,Z_1), η_1.

b) Generate a simulated vector of observed X_1 values, X_1^*, by adding the simulated residuals in a) to $(Y_1,Z_1)\hat{\alpha}$.

c) Calculate α^* by regressing X_1^* on (Y_1,Z_1).

d) Calculate imputed values, $X_0^* = (Y_0\ Z_0)\alpha^* + \eta_0^*$, where η_0^* is a N_0 element vector drawn independently with replacement from η_1 as in a).

Each loop through these four steps creates another set of imputations. Freedman's (1981) results also imply that this bootstrap imputation procedure has the same asymptotic properties as the normal imputation procedure. Small sample biases in the bootstrap can be removed by multiplying the residual vector, η_1, by $(N_1/(N_1-K))^{-1/2}$ before resampling in steps a) and d). Although this bootstrap method does not require normality, it is crucial that the residual vector, η, be homoskedastic.

When the Monte Carlo study leading to Table 1 is replicated using the above bootstrap imputation procedure, then the results are almost identical to Table 1. It would be interesting to examine the behavior of these different imputation schemes when the data generating process is not normally distributed, since that is where differences should arise. The bootstrap imputation procedure may also be easier to implement in existing statistical software packages, since it does not require explicit sampling from parametric distributions.

If the regression function, $E(X|Y,Z)$, is nonlinear, then none of the above techniques will yield proper imputations. Assuming that $E(Y|X,Z)$ is still linear and homoskedastic, the imputation methods could be modified by replacing the least squares approximation to $E(X|Y,Z)$ with some other consistent estimator. Manski (1991) gives a recent review of possible estimators. As long as imputations generated according to equation (17) asymptotically satisfy the moment conditions in equations (18), (19), (21) and (22), the resulting multiple imputations estimators should still be consistent. In practice, most models with $E(X|Y,Z)$ nonlinear will also have $E(Y|X,Z)$ nonlinear. Brownstone and Golob (1992) used multiple imputations in a model where $(X,Y|Z)$ follow a joint ordered probit distribution. Monte Carlo studies and internal consistency checks suggest that multiple imputations yields consistent inferences in their application.

One possible difficulty with all of the imputation procedures discussed above is that when (Y_0, Z_0) contains outliers relative to (Y_1, Z_1), regression predictions can be far outside the range of the observed values, X_0. In many cases this means that the imputed values are the ones with the highest leverage in the completed case estimations. Little (1988) has proposed a method, called predictive mean matching, which avoids imputing extreme values. Predictive mean matching uses the output from one of the other imputation procedures and then assigns the observed value in X_1 which is closest to the imputed value as the final imputed value. This method can introduce large biases unless the range of the observed values, X_1, includes the range of the unobserved true values, X_0.

Multiple imputations using either Little's predictive matching or the new bootstrap imputation procedure should be more robust to departures from normality than maximum likelihood methods.

4. Multiple Imputations for Nonignorable Missing Data

The previous sections of this paper have all assumed that the missing data process is ignorable, which means that, conditioned on Z, (X_1,Y_1) is a simple random sample from (X,Y). In this case the complete data estimator, $\hat{\theta}_1$, is consistent and the main issues are efficiency and consistent inference. If the missing data process is nonignorable, then $\hat{\theta}_1$ and all of the imputation procedures discussed in the previous section are inconsistent. One common approach (see Heckman, 1976) in applied econometrics is to postulate a joint model for the response probability and the regression equation (7) and then jointly estimate the model using the observed data. If the response probabilities are known, such as with deliberate choice–based sampling, then weighted least squares with weights proportional to the inverse response probabilities will yield consistent estimates using the observed data[1]. Once some method of consistently estimating $E(X_0 | Y_0,Z_0)$ is adopted, then any of the methods discussed in Section 3 can be used to generate proper multiple imputations for the missing observations. This section shows how multiple imputations can also be useful for consistently estimating the imputation models when there are non–ignorable missing data.

Although weighted regression methods are simple to use, inferences from these procedures are only valid for known fixed sampling weights. In many cases it may be possible to consistently estimate the sampling weights, but then inference procedures need to be modified to account for the estimation error in the sampling

---

[1] See DuMouchel and Duncan (1983). Note that this is just the Weighted Exogenous Sample Maximum Likelihood Estimator (Manski and Lerman, 1977) applied to the linear regression model. DuMouchel and Duncan point out that the correct covariance estimator for weighted least squares in this situation is given by $s^2(X'DX)^{-1}(X'D^2X)(X'DX)^{-1}$. Unfortunately, most weighted least squares packages use the GLS formula $s^2(X'DX)^{-1}$ which is inconsistent here.

weights. Suppose it is possible to generate multiple sets of imputed sampling weights, then consider multiple imputation estimators given in equations (1) – (3) with $\hat{\theta}_i^*$ and $\hat{\Omega}_i^*$ being the weighted regression coefficient and covariance estimators for the i^{th} set of imputed weights. Since conditional on the i^{th} set of weights, $\hat{\theta}_i^*$ and $\hat{\Omega}_i^*$ are clearly consistent, $\bar{\theta}_M$ and $\bar{\Omega}_M$ are consistent for θ and $E(Cov(\bar{\theta}_M|\text{weights}))$ (at least as $M \to \infty$). Since $E(\bar{\theta}_M|i^{th}\text{ set of weights}) = \hat{\theta}_i^*$, as M goes to infinity B_M is consistent for $Cov(E(\bar{\theta}_M|\text{weights}))$. Therefore, by equation (4), T_M is consistent for $Cov(\bar{\theta}_M)$ when both N and M go to infinity.

Rubin (1986) gives a method for estimating and multiply imputing weights for statistical file matching. Brownstone and Golob (1992) use this method to multiply impute weights needed to predict the number of commuters who would carpool to work as a function of the level of various carpooling incentives. A small Monte Carlo study established the validity of the multiple imputations inferences for this example. However, since the estimation error in the weights is very small in this application (i.e. B_M is 5% of T_M), this is not a very demanding test of the methodology.

Since the Weighted Exogenous Sample Maximum Likelihood Estimator (WESMLE, see Manski and Lerman, 1977) is a linear function of the weights, multiple imputations should yield consistent inferences for the WESMLE applied to nonlinear models. This application might prove useful in handling attrition from panel data, where the attrition probabilities (and therefore response probabilities and weights) could be estimated using pre–attrition wave data. Since most large surveys produce estimated final weights (called "post–stratification" or "non–response reweighting"), the multiply imputed WESMLE developed here should have broad applicability.

Multiple imputations can also be useful for estimating standard sample selection models. It is common to use Heckman's (1976) two–step procedure to

estimate these models. Unfortunately, it is sufficiently difficult to obtain consistent standard errors (see Lee, Maddala, and Trost, 1980) for the two–step procedure that they are rarely computed in applied work. Multiply imputing values for the Mill's Ratio yields a computationally simpler consistent variance estimate. This technique will be illustrated by a Monte Carlo study closely based on Brownstone and Englund's (1991) model of Swedish housing demand.

The standard sample selection model is given by:

$$z^* = W\alpha + \eta, \quad z = 1 \text{ if } z^*>0 \text{ and } = 0 \text{ otherwise,} \tag{28}$$

$$y = X\theta + \epsilon, \text{ observed only if } z = 1, \tag{29}$$

where $(\eta\ \epsilon)$ are bivariate normal $[0,0,1,\sigma^2,\rho]$. Therefore:

$$\text{Prob}\ (z=1) = \phi(W\alpha), \tag{30}$$

where ϕ is the standard normal cumulative distribution function, and

$$E(y|\, z=1) = X\theta - \rho\sigma\lambda(W\alpha)\ , \tag{31}$$

where the Mill's Ratio is defined as

$$\lambda = \frac{\phi(W\alpha)}{\phi(W\alpha)} \tag{32}$$

(ϕ is the standard normal density function). It is more efficient to estimate this model by maximum likelihood, but it is usually estimated estimated with a two stage procedure: 1) estimate α from the probit selection equation (30) to get $\hat{\alpha}$, then 2) estimate θ and $\rho\sigma$ by regressing y on X and $-\lambda(W\hat{\alpha})$.

Heckman (1976) shows that these two stage estimates are consistent and derives a consistent estimator for their sampling covariances. Unfortunately, these consistent covariances are rarely computed because of their complexity. However, conditional on $\hat{\alpha}$, the regression of y on X and $\lambda(W\hat{\alpha})$ is consistent with heteroskedastic residuals. If e_j is the residual corresponding to the j^{th} observation, then, conditional on $\hat{\alpha}$

$$\text{(33)} \qquad \text{Var}(e_j) = \sigma^2 - (\rho\sigma)^2\lambda(W_j\hat{\alpha})(\lambda(W_j\hat{\alpha}) + W_j\hat{\alpha}) .$$

Therefore,

$$\text{(34)} \qquad e'e / (N_1-K) + (\hat{\rho}\sigma)^2 N_1^{-1}\sum_j \lambda(W_j\hat{\alpha})(\lambda(W_j\hat{\alpha}) + W_j\hat{\alpha}) .$$

is a consistent estimator of σ^2, which can be used to get a consistent estimator, $\hat{\sigma}_j^2$, for $\text{Var}(e_j)$.

Feasible GLS estimation of equation (31) can then be implemented by regressing $y_j/\hat{\sigma}_j$ on $X_j/\hat{\sigma}_j$ and $-\lambda(W_j\hat{\alpha})/\hat{\sigma}_j$ yielding consistent estimates $\hat{\theta}$ and $\hat{\Omega}$ conditional on $\hat{\alpha}$. If multiple imputations of λ are drawn by making independent draws of α from the asymptotic normal distribution of $\hat{\alpha}$, then the same argument used previously in this section shows that the resulting multiple imputations estimators $\bar{\theta}_M$ and T_M are consistent for θ and the asymptotic covariance of $\bar{\theta}_M$ as M and N go to infinity. This multiply–imputed feasible GLS estimator is asymptotically more efficient than Heckman's 2–step estimator, and it is easier to compute than Lee, Maddala, and Trost's (1980) consistent covariance estimator for Heckman's procedure.

The practical utility of the above multiple imputations approach is illustrated using a simplified version of Brownstone and Englund's (1991) model of Swedish

housing tenure choice and "quantity" of housing demanded by owners. In the notation of equations (28) and (29), $z=1$ if the household owns a home and y represents quantity of owner–occupied housing measured by regionally–deflated assessed value. W and X contain age of household head, size of household, housing price measures, and measures of household income which are constructed to avoid endogeneity problems caused by the asymmetric tax treatment of owner and renter–occupied housing. To keep this example simple, I will only report results here for various estimates of the coefficients of the disposable income and the negative of the Mill's Ratio variables in the conditional demand equation. Table 2 gives results from applying various estimators to the same 665 observations used in Brownstone and Englund (1991), which includes 425 owners.

Table 2: Conditional Housing Demand Estimators

Estimator	Income Coefficient	$\rho\sigma$
MLE	2.26	0.27
	(0.39)	(0.069)
Heckman 2–Step	2.40	0.20
	(0.92)	(0.19)
Feasible GLS	2.40	0.15
	(0.91)	(0.15)

Note: Asymptotic standard errors in parentheses are computed using: Berndt, Hall, Hall and Hausman (1974) estimator for MLE, Lee, Maddala, and Trost (1980) estimator for 2–step, and multiple imputations described above for Feasible GLS.

The MLE appears much more efficient than the other estimators, but there does not seem to be much difference between the 2–step and Feasible GLS estimators in this example. The usual least squares standard error estimates from

the second step of either the 2–step or Feasible GLS estimator are approximately 50% of the consistent values given in Table 2, which highlights the importance of getting consistent standard errors for these estimators. This downward bias is expected from the asymptotic results in Lee, Maddala, and Trost (1980).

To guard against the possibility that the results in Table 2 are contaminated by model misspecification, a Monte Carlo study of the 2–step and Feasible GLS estimator was performed using the MLE estimates applied to equations (28) and (29) as the data generating process. The MLE itself is not included in this study because of convergence problems with some of the Monte Carlo samples. These Monte Carlo estimates, given in Table 3, can also be interpreted as parametric bootstrap estimates of the sampling variability of the two estimators.

Table 3: Monte Carlo Results for Conditional Housing Demand Estimators

	Income Coefficient		$\rho\sigma$	
Estimator	Mean	Std. Dev.	Mean	Std. Dev.
2–Step	1.91	0.78	0.27	0.16
SE(2–Step)	0.92	0.25	0.19	0.068
Feasible GLS	1.54	0.72	0.25	0.16
SE(FGLS)	0.81	0.18	0.19	0.062

Note: SE($\cdot$) represent the same consistent standard error estimators used in Table 2.

Table 3 shows the same similarity between the estimators as in Table 2, although there is some indication that the variability of the multiple imputations standard error estimator is lower than Lee, Maddala and Trost's estimator. The main difference between these estimators is computational; Lee, Maddala and Trost's estimator requires manipulation of N×K matrices, while the multiple

imputations estimator requires repetitive manipulation of K×K matrices.

5. Conclusions

Econometricians have avoided imputing values for missing data since this can lead to seriously biased inferences. Multiple imputations is a general method for consistent inferences with imputed values. This paper has modified and extended Schenker and Welsh's (1988) results to directly prove asymptotic normality of the multiple imputations point estimator and consistency of the covariance estimator for univariate and multivariate endogenous or exogenous missing data in linear regression models. Similar methods, together with linearization, should yield similar results for nonlinear models estimated by maximum likelihood or minimum distance techniques. In addition to these theoretical results and Rubin's (1987) Bayesian analysis, the Monte Carlo studies and empirical examples described here show that multiple imputations is a useful addition to applied econometricians' toolkits.

Although typically not fully efficient, multiple imputations estimators are relatively easy to compute for a wide variety of problems. When the the new bootstrap imputation methods discussed in Section 3 are used, multiple imputations are also less sensitive to distributional assumptions than parametric likelihood methods. As Rubin (1987) and Schenker, Treiman, and Weidman (1988) have pointed out, distribution of multiply-imputed public use data sets provides a new approach for communicating the accuracy of the data collected in large surveys like the PSID and CPS. This would provide much more quantitative information than the currently available imputation flags or accuracy codes.

The multiple imputations technique is no substitute for careful joint modeling of the missing data process and all variables affected by missing data. The strengths of the method are computational simplicity, flexibility, and, when bootstrap–type imputation methods are used, robustness against small–sample normality assumptions. In addition to their use in missing data problems, Section 4 also shows how multiple imputations can be used to get consistent covariance estimators for Heckman's 2–step estimator in the sample selection model and for the WESMLE with estimated weights.

References

Berndt, E., B.H. Hall, R.E. Hall, and J.A. Hausman (1974), "Estimation and Inference in Nonlinear Structural Models," *Annals of Economics and Social Measurement*, 3, 653–665.

Brownstone, D. and P. Englund (1991), "The Demand for Housing in Sweden: Equilibrium Choice of Tenure and Type of Dwelling", *Journal of Urban Economics*, 29, 267–281.

Brownstone, D. and T.F. Golob (1992), "The Effectiveness of Ridesharing Incentives: Discrete–choice Models of Commuting in Southern California," *Regional Science and Urban Economics*, forthcoming 1992.

Brownstone, D. and R. Valletta (1991), "Modeling Measurement Error Bias in Cross–Section and Longitudinal Wage Equations," UCI Department of Economics Working Paper, September, 1991.

DuMouchel, W.H. and G.J. Duncan (1983), "Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples," *Journal of the American Statistical Association*, 78, 535–543.

Freedman, D.A. (1981), "Bootstrapping Regression Models," *Annals of Statistics*, 9, 1218–1228.

Fuller, W. (1987), *Measurement Error Models*, John Wiley and Sons, New York.

Heckman, J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables, and a Simple Estimator for such Models," *Annals of Economics and Social Measurement*, 5, 475–492.

Herzog, T.N. and D.B. Rubin (1983), "Using Multiple Imputations to Handle Nonresponse in Sample Surveys," in W.G. Madow, I. Olkin, and D.B. Rubin , eds., *Incomplete Data in Sample Surveys*, V. 2, Academic Press, New York, 185–297.

Lee L–F., G.S. Maddala, and R.P. Trost (1980), "Asymptotic Covariance Matrices of Two–Stage Probit and Two–Stage Tobit Methods for Simultaneous Equations Models with Selectivity," *Econometrica*, 48, 491–504.

Li, K–H., X–L. Meng, T.E. Raghunanthan, and D.B. Rubin (1991), "Significance Levels from Repeated P–values with Multiply–Imputed Data," *Statistica Sinica*, 1, 65–92.

Little, R. J. A. (1988), "Missing–Data Adjustments in Large Surveys," *Journal of Business and Economic Statistics*, 6, 287–301.

Little, R. J. A. and D. B. Rubin (1987), *Statistical Analysis with Missing Data*, John Wiley.

Manski, C.F. (1991), "Regression," *Journal of Economic Literature*, 29, 34–50.

Manski, C.F. and S. Lerman (1977), "The Estimation of Choice Probabilities from Choice–Based Samples," *Econometrica*, 45, 1977–1988.

Press, S.J. (1982), *Applied Multivariate Analysis*, 2nd Edition, Robert E. Krieger Publishing Company, Malabar, Florida.

Rubin, D.B. (1986), "Statistical Matching Using File Concatenations with Adjusted Weights and Multiple Imputations," *Journal of Business and Economic Statistics*, 4, 87–94.

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

Ruud, P.A. (1991), "Extensions of Estimation Methods Using the EM Algorithm," *Journal of Econometrics*, 49, 305–341.

Schenker, N., D.J. Treiman, and L. Weidman (1988), "Evaluation of Multiply–Imputed Public–Use Tapes," *Proceedings of the American Statistical Association Survey Research Methods Section*, 85–92.

Schenker, N. and A. H. Welsh (1988), "Asymptotic Results for Multiple Imputation," *Annals of Statistics*, 16, 1550–1566.