The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- CompuMath Citation Index®
- Current Contents/Social and Behavioral Sciences®
- RePEc: Research Papers in Economics
- Science Citation Index Expanded (also known as SciSearch®)
- Scopus™
- Social Sciences Citation Index®

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata, Mata, NetCourse, and Stata Press are registered trademarks of StataCorp LP.

The Stata Journal (2011)
11, Number 3, pp. 403–419

A closer examination of three small-sample approximations to the multiple-imputation degrees of freedom

David A. Wagstaff
HHD Consulting Group
College of Health and Human Development
Pennsylvania State University
daw22@psu.edu

Ofer Harel
Department of Statistics
University of Connecticut

Abstract. Incomplete data is a common complication in applied research. In this study, we use simulation to compare two approaches to the multiple imputation of a continuous predictor: multiple imputation through chained equations and multivariate normal imputation. This study extends earlier work by being the first to 1) compare the small-sample approximations to the multiple-imputation degrees of freedom proposed by Barnard and Rubin (1999, *Biometrika* 86: 948–955); Lipsitz, Parzen, and Zhao (2002, *Journal of Statistical Computation and Simulation* 72: 309–318); and Reiter (2007, *Biometrika* 94: 502–508) and 2) ask if the sampling distribution of the t statistics is in fact a Student's t distribution with the specified degrees of freedom.

In addition to varying the imputation method, we varied the number of imputations ($m = 5, 10, 20, 100$) that were averaged over 500,000 replications to obtain the combined estimates and standard errors for a linear model that regressed the log price of a home on its age (years) and size (square feet) in a sample of 25 observations. Six age values were randomly set equal to missing for each replication.

As assessed by the absolute percentage and relative percentage bias, the two approaches performed similarly. The absolute bias of the regression coefficients for age and size was roughly -0.1% across the levels of m for both approaches; the absolute bias for the constant was 0.6% for the chained-equations approach and 1.0% for the multivariate normal model. The absolute biases of the standard errors for age, size, and the constant were 0.2%, 0.3%, and 1.2%, respectively. In general, the relative percentage bias was slightly smaller for the chained-equations approach. Graphical and numerical inspection of the empirical sampling distributions for the three t statistics suggested that the area from the shoulder to the tail was reasonably well approximated by a t distribution and that the small-sample approximations to the multiple-imputation degrees of freedom proposed by Barnard and Rubin and by Reiter performed satisfactorily.

Keywords: st0235, missing data, multiple imputation, small-sample degrees of freedom

1 Introduction

Missing data are present in most studies that ask individuals to report on their behavior. With early efforts to address missing data (for example, mean substitution, last observation carried forward, and single hot-deck imputation), researchers replaced each missing value with some plausible value. These early methods fell short because they did not account for the uncertainty introduced by the values that were substituted for the missing observations.

Beginning in 1976, Rubin and colleagues have championed multiple imputation (MI) as a flexible, general-purpose method for dealing with missing data (Rubin 1976, 1977, 1996; Schafer 1997, 1999; Little and Rubin 2002; Harel and Zhou 2007). With MI, researchers replace each missing value with $m > 1$ plausible values. These values are independent draws from the conditional distribution of the missing data given the observed data. Moreover, these draws are based on a parametric or semi-parametric model (that is, an imputation model) for the joint distribution, which is used to derive the conditional distribution of the missing data given the observed data.

Judging from its widespread implementation in numerous statistical programs (for example, R, SAS, SPSS, Stata), many researchers use a multivariate normal model to represent the joint distribution and calculate imputed values. In part, MI is widely used by researchers from different disciplines because Rubin proposed simple rules that researchers could use to obtain combined estimates and their standard errors, and to account for the uncertainty induced by the practice. Although Rubin used a Bayesian perspective to develop MI (Rubin 1987), he and his colleagues have shown that the obtained estimates demonstrate good frequentist properties (Little and Rubin 2002).

1.1 Rubin's rules for obtaining combined estimates and assessing their variability

"Rubin's rules" for obtaining combined estimates and standard errors advanced current practice because their use properly reflected the variability of the estimand, both within and between the m multiply imputed datasets (Rubin and Schenker 1986). Additionally, the rules allowed researchers to use standard, complete-data methods with their analysis of each multiply imputed dataset.

For a scalar estimand θ (for example, a regression coefficient), the combined estimate was the arithmetic average of the m estimates calculated with the observed and imputed data in each multiply imputed dataset ($\widehat{\theta}_i, i = 1, \ldots, m$), and the variance of the combined estimate, T, was the sum of two components: 1) the average variability of the m within-dataset variances, $\overline{W} = 1/m \sum_{i=1}^{m} W_i$, where W_i, the variance of the ith dataset, was calculated in the usual manner using the observed and imputed data, and 2) the variability among the m estimates, $B = 1/(m-1) \sum_{i=1}^{m} \left(\widehat{\theta}_i - \overline{\widehat{\theta}}\right)^2$, where $\overline{\widehat{\theta}} = 1/m \sum_{i=1}^{m} \widehat{\theta}_i$. More specifically, the total variance was $T = \overline{W} + \{(m+1)/m\} B$.

With a large sample, Rubin and Schenker argued that interval estimates for the combined estimate, $\bar{\hat{\theta}} \pm t_{\upsilon}\sqrt{T}$, could be based on Student's t distribution with υ degrees of freedom where $\upsilon = (m-1)\left[1 + \{1/(m+1)\}\left(\overline{W}/B\right)\right]^2$.

1.2 Approximations to the small-sample degrees of freedom

In deriving the rules for calculating the combined estimate and its standard error and in determining the degrees of freedom for the reference t distribution, Rubin and Schenker (1986) assumed that the size of the complete dataset was large enough to use large-sample methods that effectively set the t statistic's degrees of freedom equal to infinity. Specifically, they assumed that there were an infinite number of observations in the complete dataset. Soon, researchers reported that when the complete-data degrees of freedom, υ_{com}, was small and there was a modest proportion of missing data, Rubin and Schenker's approximation for the MI degrees of freedom, υ_m, could be many times larger than the complete-data degrees of freedom (the available degrees of freedom in the absence of any missing data).

To address the problem with Rubin and Schenker's approximation, Barnard and Rubin (1999) proposed a small-sample adjustment for the MI degrees of freedom that would always be less than or equal to the complete-data degrees of freedom. Then they conducted a simulation study to determine how well the empirical distribution of the combined estimate of the slope of a simple linear regression model followed a t distribution with their adjusted degrees of freedom. For their study, they generated bivariate normal data and varied five factors: the correlation between y and x ($\rho = 0.5, 0.8$); the sample size ($N = 10,\ 20,\ 30$); the number of imputations ($m = 3,\ 5,\ 10$); the percentage of missingness ($\varpi = 10,\ 20,\ 30$); and the slope of their missing-data function ($\eta = -4,\ 0,\ 4$). Using 1,000 replications for each of the 162 cells of their study design, Barnard and Rubin found that their proposed modification exhibited better coverage than did Rubin and Schenker's large-sample approximation for all conditions.

Lipsitz, Parzen, and Zhao (2002) proposed another small-sample approximation to the MI degrees of freedom. They submitted their article in 2001 and did not cite Barnard and Rubin's (1999) study or compare their proposed approximation to Barnard and Rubin's approximation. They simply noted that Rubin and Schenker's (1986) approximation might be inaccurate when the sample size was small. Specifically, Lipsitz, Parzen, and Zhao (LPZ) proposed the following small-sample approximation to the multiply imputed degrees of freedom for a scalar estimand:

$$\upsilon_{\text{LPZ}} = \frac{\left[\overline{W} + \{(m+1)/m\}\, B\right]^2}{\left\{\overline{W}^2/(N-1)\right\} + \{(m+1)m\}^2\left\{B^2/(m-1)\right\}}$$

which can be rewritten as

$$v_{\text{LPZ}} = \frac{\left[(\overline{W}/B) + \{(m+1)/m\}\right]^2}{\left[\{1/(N-1)\}\left(\overline{W}/B\right)^2 + \{(m+1)/m\}^2\{1/(m-1)\}\right]}$$

This reexpression makes it easier to see that the Lipsitz, Parzen, and Zhao approximation approaches the nominal degrees of freedom for the test that a population mean is equal to zero as m increases for any two values of the variance components. Moreover, by reexpressing their approximation in terms of $\overline{W}/B$, Lipsitz, Parzen, and Zhao make it easier to appreciate how it differs from Rubin and Schenker's approximation.

In deriving their approximation, Lipsitz, Parzen, and Zhao (2002) started with the same expression that Rubin and Schenker (1986) had used for the variance of the combined estimate. However, they assumed that each multiply imputed sample consisted of a finite number of observations, N. In addition, they assumed that the within and between mean squares were independently distributed on $N-1$ and $m-1$ degrees of freedom, respectively. As a practical consequence, they noted that their approximation to the degrees of freedom was always less than or equal to Rubin and Schenker's. Thus confidence intervals based on their approximation would be larger than those based on Rubin and Schenker's approximation.

To evaluate their approximation, Lipsitz, Parzen, and Zhao (2002) fit a multiple-regression model to 25 observations taken from housing market data published in the Dallas Morning News in 1990. Specifically, they regressed the natural logarithm of the house's selling price (in hundreds of thousands of dollars) on its size (in units of a thousand square feet of heated floor space) and age (in years). With the exception of age, which was not observed for 6 of the 25 houses, all the remaining values were observed. Lipsitz, Parzen, and Zhao used a multivariate normal imputation model as implemented in Schafer's (1997) S-Plus program to impute the missing values with $m = 5$ and $m = 20$. When they compared their approximation with Rubin and Schenker's (1986) large-sample approximation, they found that their estimates of the degrees of freedom for the three regression coefficients were much smaller than Rubin and Schenker's and that the latter's large-sample approximation essentially used the normal distribution as the reference distribution whether $m = 5$ or $m = 20$.

Recently, Reiter (2007; Marchenko and Reiter 2009) proposed another small-sample approximation to the MI degrees of freedom. Like Barnard and Rubin (1999), Reiter noted that the degrees of freedom suggested by Rubin and Schenker (1986) could be larger than the complete-data degrees of freedom. However, Reiter approached the problem from the perspective of the researcher who wants to conduct a multicomponent test where the combined estimate was a vector and the reference distribution was an F distribution. Reiter noted that one of the assumptions that Rubin and Schenker used to derive their expression for the degrees of freedom was that the sample size was infinite. Moreover, as he was addressing multicomponent significance tests, Reiter noted that the degrees of freedom suggested by Li, Raghunathan, and Rubin (1991) and by Meng and Rubin (1992) could exceed the degrees of freedom that would be used if there were no missing data. Consequently, Reiter proposed the following approximation:

$$
\begin{aligned}
v_f &= 4 + 1/z \\
z &= \frac{1}{v^*_{\text{com}} - 4(1+a)} + \frac{1}{t-4}\left[\frac{a^2\left\{v^*_{\text{com}} - 2\left(1+a\right)\right\}}{\left(1+a\right)^2\left\{v^*_{\text{com}} - 4\left(1+a\right)\right\}}\right] \\
&\quad + \frac{1}{t-4}\left[\frac{8a^2\left\{v^*_{\text{com}} - 2\left(1+a\right)\right\}}{\left(1+a\right)\left\{v^*_{\text{com}} - 4\left(1+a\right)\right\}^2} + \frac{4a^2}{\left(1+a\right)\left\{v^*_{\text{com}} - 4\left(1+a\right)\right\}}\right] \\
&\quad + \frac{1}{t-4}\left[\frac{4a^2}{\left\{v^*_{\text{com}} - 4\left(1+a\right)\right\}\left\{v^*_{\text{com}} - 2\left(1+a\right)\right\}} + \frac{16a^2\left\{v^*_{\text{com}} - 2\left(1+a\right)\right\}}{\left\{v^*_{\text{com}} - 4\left(1+a\right)\right\}^3}\right] \\
&\quad + \frac{1}{t-4}\left[\frac{8a^2}{\left\{v^*_{\text{com}} - 4\left(1+a\right)\right\}^2}\right]
\end{aligned}
$$

where $v^*_{\text{com}} = \left\{\left(v_{\text{com}} + 1\right) / \left(v_{\text{com}} + 3\right)\right\} v_{\text{com}}$; $a = r_m t/(t-2)$; $t = k(m-1)$; and $r_m = (1 + 1/m)tr\left(B\overline{W}^{-1}\right)/k$. In the above, m is the number of imputations; k is the number of components that were tested; v_{com} is the complete-data degrees of freedom; B is the between-variance component; and $\overline{W}$ is the average within-variance component. Reiter noted that his approximation used some results from Barnard and Rubin (1999). However, his approximation would always be less than or equal to the complete-data degrees of freedom for samples of modest size and equal to the standard degrees of freedom for samples of infinite size. Reiter did not cite the Lipsitz, Parzen, and Zhao (2002) study.

To compare the performance of his approximation, Reiter conducted a simulation study. He obtained 10,000 replications each for two conditions that varied the sample size and the number of assessed coefficients ($N = 50$ and $k = 4$; $N = 200$ and $k = 9$). The data were generated as a k-variate multivariate normal sample with a mean of 0 and a variance–covariance matrix that exhibited compound symmetry ($\rho = 0.5$); and the response was an independently generated standard normal variate. Missing data were generated by randomly deleting 10% of all values in the dataset. Reiter used SAS' Proc MI and its multivariate normal model to obtain five multiply imputed datasets. Reiter reported that his small-sample approximation was calibrated better than was the standard approximation of Li, Raghunathan, and Rubin (1991) and Meng and Rubin (1992) for multicomponent significance tests with multiply imputed data.

2 The present study

In the present study, we had three objectives. First, we sought to fill a gap in the literature by comparing the three small-sample approximations to the degrees of freedom for the multiply imputed data that have been proposed: Barnard and Rubin's (1999) approximation; Lipsitz, Parzen, and Zhao's (2002) approximation; and Reiter's (2007) approximation. Second, we compared these approximations when the imputations were

calculated using Stata's multivariate normal imputation model (version 11, StataCorp 2009) and Royston's implementation of the MIs through the chained-equations approach (Royston 2004, 2005, 2007; Royston, Carlin, and White 2009; White, Royston, and Wood 2011). Although research suggests that the two approaches produce comparable results when all the conditional imputation models are linear regressions (van Buuren 2007), researchers should conduct studies that compare the different imputation procedures with different data structures and report their findings and experiences with the implementing software. Third, we compared the three small-sample approximations when a multiple regression model is fit to data collected from a small sample, and researchers would use Student's t test to assess the null hypothesis that a population regression coefficient was equal to 0.

We initiated our simulation study by generating multivariate normal data that were consistent with a complete-data version of the housing data that Lipsitz, Parzen, and Zhao (2002) had used to study the relationship between three variables reported for 25 homes in the Dallas, Texas, housing market: the price (in dollars), the age (in years), and the size of the house (in square feet). Six homes had missing data on age.

To obtain a complete-data version of Lipsitz, Parzen, and Zhao's (2002) dataset, we used Royston's (2004, 2005, 2007) missing-data program, `ice`, and randomly selected an imputed dataset as the starting point for identifying population parameters (see table 1). We then used Stata's `drawnorm` program to generate 25 multivariate normal observations and randomly set six age values equal to missing. We used a missing completely at random (MCAR) missing-data mechanism for two reasons. First, we wanted to focus on the empirical sampling distribution of the combined estimator, its estimated standard error, and the corresponding Student's t statistic. Second, we wanted to minimize the impact of both the missing-data mechanism and questions regarding specification of the imputation model. Other researchers (for example, Graham, Olchowski, and Gilreath [2007] and van Buuren [2010]) have also deemphasized the missing-data mechanism so that they could compare different imputation methods in a straightforward manner.

Table 1. Summary statistics for the multivariate normal data generator, $N = 25$

Population means

$$\begin{bmatrix} \text{log(price)} \\ \text{age(yrs)} \\ \text{size(thousands)} \end{bmatrix} = \begin{bmatrix} 11.41948 \\ 6.120000 \\ 1.874800 \end{bmatrix}$$

Population standard deviations

$$\begin{bmatrix} 0.2002466 \\ 2.1470910 \\ 0.4002216 \end{bmatrix}$$

Population correlation matrix

$$\begin{pmatrix} 1.0000 & -0.0435 & 0.7564 \\ -0.0435 & 1.0000 & -0.2469 \\ 0.7564 & -0.2469 & 1.0000 \end{pmatrix}$$

Population regression model and standard errors

log(price)	=	β_0+	$\beta_1$age+	$\beta_2$size + residual
	=	10.58754+	0.01422age+	0.39730size + residual
		(0.17203)	(0.01307)	(0.07014)

Our simulation design varied two factors: the imputation model (`mvn`, `ice`) and the number of imputations $(5, 10, 20, 100)$. We conducted 500,000 replications for each cell of our design. We chose this number because we believed that it would provide reasonable estimates of the sampling distribution's mean, variance, and percentiles (for example, the 97.5th percentile).

We used the data reported by Lipsitz, Parzen, and Zhao (2002) for two reasons. First, these observations have been used previously with a multivariate normal imputation model to compare the performance of the small-sample approximations proposed by Barnard and Rubin (1999) and by Lipsitz, Parzen, and Zhao (2002). Second, the dataset was small enough to be used in a simulation study where we could obtain, store in memory, and manipulate as many as 100 multiply imputed datasets, and repeat these steps 500,000 times. The choice of the four levels for m was based in part on the early MI literature, which indicated that researchers only needed a "few" imputations to achieve reasonable efficiency. Initially, a few meant 3 to 5 (Allison 2000, 2003; Schafer and Olsen 1998). However, as researchers gained more experience with MI methods, a few became 10 (Schafer 1999), then 20 (Schafer and Graham 2002), and then 40 or as many as 100 (Graham, Olchowski, and Gilreath 2007).

Although researchers are reevaluating the number of imputations that they need to use (Hershberger and Fisher 2003; Royston 2004; von Hippel 2005; Graham, Olchowski, and Gilreath 2007; Harel 2007; Bodner 2008; White, Royston, and Wood 2011), three of the four levels used in the present study ($m = 5, 10, 20$) reflect the number of imputations that are reported most often. The present simulation was conducted on a Dell Optiplex computer with a 3.60 GHz Pentium 4 CPU and 2 GB RAM that ran on the Windows XP Professional (SP3) operating system.

3 Simulation results

Before we present our results, we take note of the data-generation process and the graphical and numerical procedures that we used to assess our empirical sampling distributions. We conducted five training runs. On each run, we used Stata's `drawnorm` program to generate 500,000 multivariate normal datasets ($N = 25$ observations) with the inputs presented in table 1 and regressed log(price) on the house's age and size. For each dataset, we wrote the three estimated regression coefficients and standard errors to a Stata dataset and calculated t statistics corresponding to the null hypothesis that a regression coefficient equaled 0. For each empirical sampling distribution, we used Stata's `kdensity` command to obtain a density plot. We overlaid the three plots of the sampling distributions for the regression coefficients with that of a normal distribution; we overlaid three plots of the sampling distribution of our t statistics with that of a Student's t distribution on 22 degrees of freedom for the three t distributions. We used Stata's quantile plots (specifically, `qnorm` and `qqplot`) to visually inspect the central and tail-area behavior of the various sampling distributions. We expected Stata's programs to perform their functions well. We simply took these steps to accustom our eyes to distinguish atypical from typical departures from the underlying theoretical models.

We also used five runs to assess how well Stata's multivariate normal generator returned samples that had a structure consistent with our target parameters. Each run consisted of 500,000 replications; and each replication fit the regression model shown in table 1 to 25 multivariate normal observations generated according to the values shown in table 1. Koehler, Brown, and Haneuse (2009) noted that few researchers have attempted to assess or report the Monte Carlo error associated with their simulation and they proposed that researchers use the standard deviation of the Monte Carlo estimator across the repetitions of the simulation. In keeping with their proposal, we note that the coefficients of variation, $100 \times (\widehat{\sigma}/\widehat{\mu})$, for the regression coefficients for `age`, `size`, and the `constant`, were 0.2%, 0.0%, and 0.0%, respectively, and that the coefficients of variation for their respective standard errors were 0.0%, 0.0%, and 0.0%.

These five runs with Stata's `drawnorm` program also showed that 7 of the 15 estimates of the t distribution's degrees of freedom obtained as $\widehat{v} = 2\widehat{\sigma}^2/\left(\widehat{\sigma}^2 - 1\right)$ were less than the nominal degrees of freedom of 22, even when the regression coefficients and standard errors were estimated with reasonable precision. The relative percentage error ranged from -5.5% to 3.6% (mean $= -0.4\%$, standard deviation $= 3.2\%$).

Table 2 presents the multiply imputed estimates of the regression coefficients for `age`, `size`, and the `constant` and their standard errors by imputation approach (the MI through chained-equations approach, `ice`, and the multivariate normal imputation model, `mvn`) and number of imputations ($m = 5$, 10, 20, and 100). Again each cell entry is based on 500,000 repetitions. As in the Lipsitz, Parzen, and Zhao (2002) study, age was the only variable that had missing data. Specifically, six values were randomly set equal to missing.

The cell entries presented in table 2 suggest that the performance of the two MI approaches was similar for the fit of the linear regression model to data from a small

sample ($N = 25$) when the missing-data mechanism is MCAR, and only one of two predictors had missing data. The cell means for the standard errors suggest that estimation did not improve appreciably when the total imputation variance was based on 100 imputations as opposed to 5 imputations. Both imputation methods overestimated the standard errors for all three regression coefficients to the same extent. It is understandable that the standard errors would be overestimated. The target values were based on the inputs given to Stata's multivariate normal generator, which were not subject to the uncertainty associated with repeatedly sampling 25 observations or the uncertainty associated with repeatedly sampling observations from a posterior distribution.

Table 2. Combined estimates of the regression coefficients and standard errors by imputation method and number of imputations ($T = 500,000$)

Variable	MI procedure	Target	Number of imputations 5	10	20	100
			MI combined regression coefficient			
`age`	`ice`	0.01422	0.01352	0.01357	0.01356	0.01354
	`mvn`		0.01307	0.01309	0.01310	0.01309
`size`	`ice`	0.39730	0.39644	0.39649	0.39637	0.39641
	`mvn`		0.39578	0.39589	0.39582	0.39596
`constant`	`ice`	10.58754	10.59346	10.59305	10.59340	10.59346
	`mvn`		10.59748	10.59715	10.59730	10.59699
			MI standard error			
`age`	`ice`	0.01307	0.01501	0.01489	0.01481	0.01476
	`mvn`		0.01499	0.01486	0.01479	0.01472
`size`	`ice`	0.07014	0.07292	0.07282	0.07267	0.07264
	`mvn`		0.07320	0.07298	0.07290	0.07277
`constant`	`ice`	0.17203	0.18591	0.18517	0.18447	0.18420
	`mvn`		0.18632	0.18525	0.18472	0.18415

Figure 1 presents the relative percentage bias associated with each imputation method across the four levels of m. Each data point was based on 500,000 replications. The figure indicates that the average of the estimates obtained with `ice` was consistently closer to the target value for the regression coefficients for `age` and `size` than was the corresponding average obtained with the multivariate normal imputation model. Additionally, the relative percentage bias suggests that differences between the two imputation approaches for a continuous response that is normally distributed may decrease as the magnitude of the estimated coefficient increases. Both the relative percentage bias and absolute percentage bias were calculated using the values displayed in table 2.

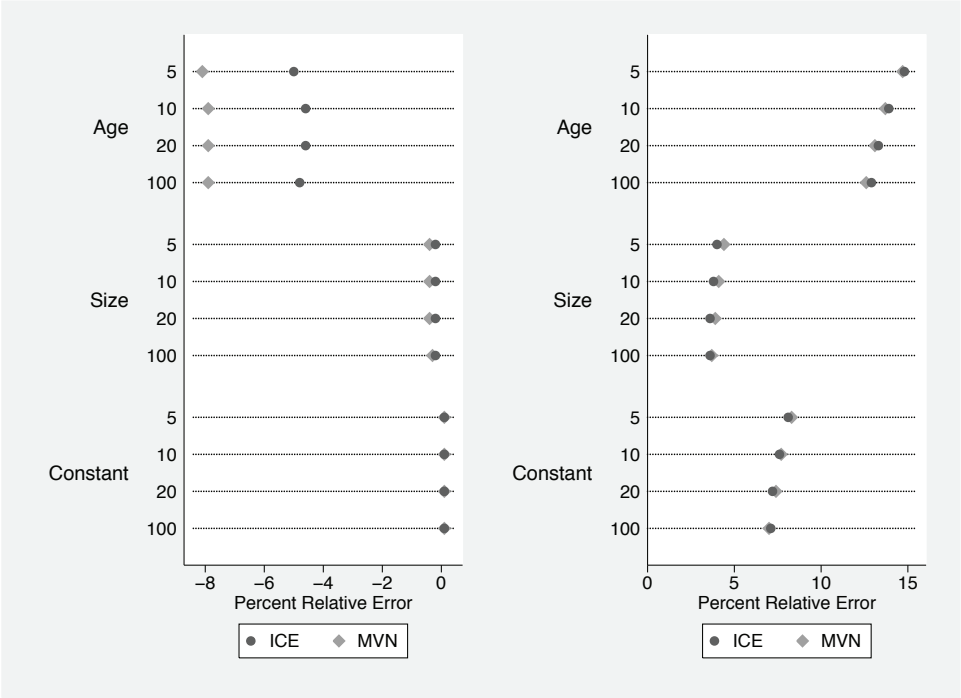

Figure 1. Relative percentage error for the regression coefficients (left panel) and their standard errors (right panel) by number of imputations and imputation method

Table 3 presents the mean of the large-sample approximation to the MI degrees of freedom proposed by Rubin and Schenker (1986) and the mean of the three small-sample approximations proposed by Barnard and Rubin (1999); Lipsitz, Parzen, and Zhao (2002); and Reiter (2007). Each mean was averaged across 500,000 repetitions that were conducted with `ice` and with Stata's `mvn` imputation program, with $m = 5$, 10, 20, and 100 imputations. With each small-sample approximation, the mean increased as the number of imputations increased from $m = 5$ to $m = 100$. This increase was larger for the Lipsitz, Parzen, and Zhao approximation than it was for Barnard and Rubin's approximation, and the increase observed with Barnard and Rubin's approximation was larger than that observed for Reiter's approximation. Given that the combined estimates for the regression coefficients were relatively stable as m increased from 5 to 100, the increase observed in the degrees of freedom for each small-sample approximation may signal that too few imputations were used to obtain a stable estimate of a standard error.

Table 3. Means for the four approximations to the multiply imputed degrees of freedom by number of imputations and imputation method

		ice				mvn			
Variable	m	RS	BR	LPZ	R	RS	BR	LPZ	R
age	5	3892.8	13.0	22.5	18.4	1808.1	12.4	22.0	18.1
	10	617.3	14.2	28.6	19.4	474.6	13.7	28.7	19.2
	20	895.4	14.9	33.0	19.7	712.2	14.4	33.8	19.6
	100	3876.4	15.5	37.4	19.8	3151.2	15.1	39.3	19.7
size	5	24700000.0	18.2	23.8	20.0	5412033.0	18.1	23.8	19.9
	10	635068.9	18.6	24.7	20.1	309759.5	18.5	24.9	20.1
	20	356568.9	18.8	25.0	20.1	295088.9	18.7	25.2	20.1
	100	1150215.0	19.0	25.2	20.1	856355.1	18.9	25.5	20.1
constant	5	42794.1	15.9	24.1	19.5	34472.9	15.6	24.0	19.4
	10	6277.3	16.8	27.0	19.9	4709.4	16.5	27.3	19.8
	20	7786.3	17.2	28.5	20.0	6070.9	17.0	29.0	19.9
	100	30550.6	17.6	29.4	20.0	23453.0	17.3	30.2	20.0

Notes: Each cell mean is based on 500,000 repetitions.
RS—Rubin and Schenker's (1986) large-sample approximation
BR—Barnard and Rubin's (1999) small-sample approximation
LPZ—Lipsitz, Parzen, and Zhao's (2002) small-sample approximation
R—Reiter's (2007) small-sample approximation

The respective mean degrees of freedom for the three small-sample approximations were essentially comparable across the two imputation procedures. The mean degrees of freedom for Lipsitz, Parzen, and Zhao's (2002) small-sample approximation exceeded the complete-data degrees of freedom for 11 of the 12 cells of the study design. In table II of their article, Lipsitz, Parzen, and Zhao reported that their approximation yielded values greater than the complete-data degrees of freedom for the two levels of m that they considered ($m = 5$ and $m = 20$). However, Lipsitz, Parzen, and Zhao simply sought to show that their approximation was better calibrated than was Rubin and Schenker's (1986) approximation.

Finally, the mean degrees of freedom for the three small-sample approximations exhibited the same order for each coefficient and level of m. Specifically, Barnard and Rubin's (1999) small-sample approximation for the degrees of freedom was always smaller than the remaining two approximations, and the mean degrees of freedom for Reiter's (2007) approximation was always less than the mean observed for Lipsitz, Parzen, and Zhao's (2002) approximation. Moreover, the mean degrees of freedom for Reiter's approximation was noticeably less variable across the coefficients and imputations than were the means of the remaining two small-sample approximations. With the chained-equations approach, the standard deviation of the 12 mean degrees of freedom for Reiter's approximation was 0.49; when the multivariate normal model was used to obtain the imputations, it was 0.57. Both of these standard deviations were smaller than the

respective standard deviations of the 12 means for Lipsitz, Parzen, and Zhao's approximation (4.29 and 4.89) and Barnard and Rubin's approximation (1.95 and 2.13).

4 Discussion

In this study, we used simulation to extend previous studies that have compared the performance of MI through the chained-equations approach and through imputation with a multivariate normal model in two directions. First, we compared the three small-sample approximations to the multiply imputed degrees of freedom that have been proposed by Barnard and Rubin (1999), Lipsitz, Parzen, and Zhao (2002), and Reiter (2007). Second, we may be the first to use sampling distributions with 500,000 observations to ask if a Student's t distribution is the appropriate reference distribution with the specified degrees of freedom. The dataset was modeled on the 25 observations that Lipsitz, Parzen, and Zhao had considered in their simulation study. Specifically, the natural logarithm of the house's price was described as a linear function of its `age` (years) and `size` (square feet). This dataset was chosen because it was quite small and thus could be expected to provide a moderately extreme test of the performance of each small-sample approximation. As with the Lipsitz, Parzen, and Zhao study, six age values were randomly set equal to missing prior to imputing the data.

We found that the performance of the two MI approaches was quite similar for the fit of a relatively straightforward linear regression model to data from a small sample when the missing data were MCAR and only one of two predictors had any missing data. More specifically, we found that the absolute bias of the regression coefficients for `age` and `size` was roughly -0.1% across the four levels of m ($m = 5, 10, 20, 100$) for both approaches; it was 0.6% across the levels of m for the combined estimate of the constant for the chained-equations approach, and it was 1.0% for the multivariate normal model. Moreover, the absolute bias of the corresponding standard errors was only slightly larger for `age` (0.2%), `size` (0.3%), and the `constant` (1.2%); and the relative percentage bias was slightly smaller for the chained-equations approach than it was for imputation via the multivariate normal model.

Both imputation methods overestimated the standard errors for all three regression coefficients to the same extent. The cell means for the estimated standard errors were not appreciably closer to the corresponding target value when the total imputation variance was based on 100 imputations as opposed to 5 imputations. This suggests that the components-of-variance argument used to derive the total variance of a combined estimate may not perform particularly well when the sample size is as small as 25 and there are few variables that the researcher can use in an imputation model.

We also found that the means of the degrees of freedom for the three small-sample approximations were essentially comparable across the two MI procedures. However, the mean degrees of freedom for Lipsitz, Parzen, and Zhao's (2002) small-sample approximation exceeded the complete-data degrees of freedom for 11 of the 12 cells of our study design. That their approximation could yield values greater than the complete-data degrees of freedom was also reported by Lipsitz, Parzen, and Zhao, who conducted

a simulation with $m = 5$ and $m = 20$. The present study extends their results by considering two additional levels of m (10, 100), an additional imputation approach (the chained-equations approach), and a greater number of repetitions (500,000 versus 2,000). When we used the mean degrees of freedom to characterize the performance of each approximation, we found that Barnard and Rubin's (1999) small-sample approximation was always smaller than either of the remaining two approximations, and that the mean degrees of freedom for Reiter's (2007) approximation was always less than that observed for Lipsitz, Parzen, and Zhao's approximation. Additionally, we found that the mean degrees of freedom for Reiter's approximation was noticeably less variable across the coefficients and imputations than were the means of the remaining two small-sample approximations.

5 Study limitations

The present study was intentionally restricted to a linear regression model that described the price of a home in terms of its age and size. This model was fit to 25 observations generated with Stata's multivariate normal random-number generator, and an MCAR missing-data mechanism was imposed, setting six age values equal to missing. Then only two specific approaches to imputation—MI through chained equations and the multivariate normal imputation model—were studied. Conclusions cannot be generalized readily to different regression models with different missing-data mechanisms and missing-data patterns, to larger sample sizes, or to different approaches to MI without further study.

6 Conclusions

The present findings suggest that the performance of the chained-equations approach was comparable to that of the multivariate normal model for the imputation of missing data on a continuous predictor in a multiple linear regression model fit to a sample of 25 observations. The regression estimates and standard errors obtained with the chained-equations approach were similar to those obtained with a multivariate normal imputation model. This finding is consistent with earlier studies that compared the two approaches (Lee and Carlin 2010; van Buuren et al. 2006).

What distinguishes the present study from these earlier studies is its examination of the empirical sampling distribution of the t statistics corresponding to the estimated regression coefficients. Specifically, the present study sought to determine if the distributions could be described by a Student's t with the small-sample approximations proposed by Barnard and Rubin (1999), Lipsitz, Parzen, and Zhao (2002), or Reiter (2007).

As Lipsitz, Parzen, and Zhao (2002) reported, their small-sample approximation to the MI degrees of freedom can exceed the nominal degrees of freedom that researchers would use with complete data. Because we used a dataset modeled on the 25 observations studied by Lipsitz, Parzen, and Zhao, we too found that their approximation

could exceed the nominal degrees of freedom. As a result, that approximation should not be used when the sample size is small.

Of the remaining two approximations, Reiter's (2007) approximation might be preferred because it is not an average of two estimates—one of which, Rubin and Schenker's (1986), was derived using an assumption about the sample size that is unlikely to be met in practice. The primary disadvantage of Reiter's approximation would be the difficulty that many researchers would have demonstrating or explaining how it was derived. Neither Reiter's nor Barnard and Rubin's (1999) small-sample approximation to the degrees of freedom exhibited excellent performance under either approach to imputation when m equaled 100. However, any disappointment should be tempered by the fact that few studies actually look at the empirical sampling distribution of the statistics that researchers use to construct interval estimates or test hypotheses.

Finally, we note that Stata provides its users with the tools they need to extend the present findings to regression problems with a larger number of predictors, larger and different variance–covariance structures, larger and more numerous sample sizes, and indeed different response types. Although these simulation studies are time-intensive when implemented on one computer, they are straightforward to conduct and could be conducted by a supervised class of students who want to experience the kinds of problems that some statisticians and applied researchers seek to answer via computer simulation.

7 Acknowledgments

This project was partially supported by award number K01MH087219 from the National Institute of Mental Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.

8 References

Allison, P. D. 2000. Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research* 28: 301–309.

———. 2003. Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology* 112: 545–557.

Barnard, J., and D. B. Rubin. 1999. Small-sample degrees of freedom with multiple imputation. *Biometrika* 86: 948–955.

Bodner, T. E. 2008. What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal* 15: 651–675.

Graham, J. W., A. E. Olchowski, and T. D. Gilreath. 2007. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science* 8: 206–213.

Harel, O. 2007. Inferences on missing information under multiple imputation and two-stage multiple imputation. *Statistical Methodology* 4: 75–89.

Harel, O., and X.-H. Zhou. 2007. Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine* 26: 3057–3077.

Hershberger, S. L., and D. G. Fisher. 2003. A note on determining the number of imputations for missing data. *Structural Equation Modeling* 10: 648–650.

Koehler, E., E. Brown, and S. J.-P. A. Haneuse. 2009. On the assessment of Monte Carlo error in simulation-based statistical analyses. *American Statistician* 63: 155–162.

Lee, K. J., and J. B. Carlin. 2010. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *American Journal of Epidemiology* 171: 624–632.

Li, K.-H., T. E. Raghunathan, and D. B. Rubin. 1991. Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association* 86: 1065–1073.

Lipsitz, S. R., M. Parzen, and L. P. Zhao. 2002. A degrees-of-freedom approximation in multiple imputation. *Journal of Statistical Computation and Simulation* 72: 309–318.

Little, R. J. A., and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: Wiley.

Marchenko, Y. V., and J. R. Reiter. 2009. Improved degrees of freedom for multivariate significance tests obtained from multiply imputed, small-sample data. *Stata Journal* 9: 388–397.

Meng, X. L., and D. B. Rubin. 1992. Performing likelihood ratio tests with multiply-imputed data sets. *Biometrika* 79: 103–111.

Reiter, J. P. 2007. Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. *Biometrika* 94: 502–508.

Royston, P. 2004. Multiple imputation of missing values. *Stata Journal* 4: 227–241.

———. 2005. Multiple imputation of missing values: Update of ice. *Stata Journal* 5: 527–536.

———. 2007. Multiple imputation of missing values: Further update of ice, with an emphasis on interval censoring. *Stata Journal* 7: 445–464.

Royston, P., J. B. Carlin, and I. R. White. 2009. Multiple imputation of missing values: New features for mim. *Stata Journal* 9: 252–264.

Rubin, D. B. 1976. Inference and missing data. *Biometrika* 63: 581–592.

———. 1977. Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association* 72: 538–543.

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

———. 1996. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 91: 473–489.

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

Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall/CRC.

———. 1999. Multiple imputation: A primer. *Statistical Methods in Medical Research* 8: 3–15.

Schafer, J. L., and J. W. Graham. 2002. Missing data: Our view of the state of the art. *Psychological Methods* 7: 147–177.

Schafer, J. L., and M. K. Olsen. 1998. Multiple imputation for multivariate missing data problems: A data analyst's perspective. *Multivariate Behavioral Research* 33: 545–571.

StataCorp. 2009. *Stata: Release 11*. Statistical Software. College Station, TX: StataCorp LP.

van Buuren, S. 2007. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research* 16: 219–242.

———. 2010. Item imputation without specifying scale structure. *Methodology* 6: 31–36.

van Buuren, S., J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin. 2006. Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation* 76: 1049–1064.

von Hippel, P. T. 2005. How many imputations are needed? A comment on Hershberger and Fisher (2003). *Structural Equation Modeling* 12: 334–335.

White, I. R., P. Royston, and A. M. Wood. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* 30: 377–399.

About the authors

David A. Wagstaff is a research technologist with the HHD Consulting Group in the College of Health and Human Development, Pennsylvania State University. He holds a PhD in psychology from the University of Massachusetts–Amherst and has worked on interventions to prevent smoking, repeat pregnancy, sexually transmitted infections, and substance use. His current research interests focus on the analysis of dependent data.

Ofer Harel is an associate professor in the Department of Statistics and a principal investigator in the Center for Health, Intervention, and Prevention, University of Connecticut. He holds a PhD in statistics from Pennsylvania State University. His research interests focus on the analysis of incomplete data and confidentiality issues related to Alzheimer's, diabetes, nutrition, HIV/AIDS, and alcohol and drug abuse prevention.