Statistica Neerlandica (2007)
doi:10.1111/j.1467-9574.2007.00377.x

Imputing continuous data under some non-Gaussian distributions

Hakan Demirtas* and Donald Hedeker†

Division of Epidemiology and Biostatistics (MC923), University of Illinois at Chicago, 1603 West Taylor Street, Chicago, IL, 60612, USA

There has been a growing interest regarding generalized classes of distributions in statistical theory and practice because of their flexibility in model formation. Multiple imputation under such distributions that span a broader area in the symmetry–kurtosis plane appears to have the potential of better capturing real incomplete data trends. In this article, we impute continuous univariate data that exhibit varying characteristics under two well-known distributions, assess the extent to which this procedure works properly, make comparisons with normal imputation models in terms of commonly accepted bias and precision measures, and discuss possible generalizations to the multivariate case and to larger families of distributions.

Keywords and Phrases: multiple imputation; normality; symmetry; skewness; kurtosis.

1 Introduction

The normality assumption is unequivocally one of the most extensively studied phenomena in statistics. Although real data rarely conform with normality, it has been regarded as a mathematical convenience for inferential purposes because of its well-understood distributional properties. Despite its critical importance, it only represents a single point in the skewness–elongation plane; and general classes of continuous distributions that span a broader spectrum in terms of symmetry and peakedness behavior (Genton, 2004) have received increased interest among statisticians. In this article, we describe multiple imputation (MI) under two univariate non-Gaussian distributions that can accommodate a wider range of distributional features. We believe that the ideas presented here could serve as a potential building block for performing MI in a multivariate setting under more general classes of densities (e.g. Tukey's classes, the Burr family, the Johnson family, the Pearson family, generalized lambda and beta families, Fleishman polynomials) that include many standard distributions as exact or approximate special cases (Burr, 1942; Johnson, 1949; Fleishman, 1978; Parrish, 1990; Morganthaler and Tukey, 2000).

*demirtas@uic.edu
†hedeker@uic.edu

Published by Blackwell Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

Multiple imputation under the normality assumption has emerged as a frequently used model-based approach in the last two decades. MI is a stochastic simulation technique that involves filling in missing data with $m > 1$ plausible values through a predictive distribution (Little and Rubin, 2002; Rubin, 2004). Once m versions of the completed datasets are obtained, one can proceed with analyzing them with standard complete-data methods, and consolidating the results into a single inferential summary. As a result, with MI, uncertainty due to missing data is formally taken into account in the modeling process. Other key advantages of MI are reviewed by Rubin (1996, 2004) and Schafer (1997, 1999). Additionally, methods, illustrative applications and implementations in various software packages have been increasingly described (Van Buuren, Boshuizen and Knook, 1999; Van Buuren and Oudshoom, 2000; Belin *et al.*, 2000; Schimert *et al.*, 2001; Demirtas and Schafer, 2003; Demirtas, 2005a,b; Demirtas and Hedeker, 2007; Demirtas *et al.*, 2007a). (For an extensive bibliography, see Rubin, 1996 and for a software review, see Horton and Lipsitz, 2001.) The fundamental step in parametric MI is filling in the missing data by drawing from the conditional distribution of the missing data given the observed data which usually entails positing a model for the data and using it to derive this conditional distribution. For continuous data, multivariate normality among the variables has been perceived as a natural assumption because the conditional distribution of the missing data given the observed data is then also multivariate normal. Recently, moving the practice of MI from normality to more general classes of densities has begun to receive attention (Liu, 1995; He and Raghunathan, 2006).

Given the restrictive nature of the normality assumption, employing distributions that span a wider range of symmetry-peakedness behavior in the imputation process may provide a reasonable way to handle non-Gaussian continuous data. In this regard, the beta and Weibull densities are sensible alternatives because they can take a variety of distributional shapes depending on the choice of parameter values. Here, we explore the relative advantages of conducting imputation inferences under these more flexible densities via a limited simulation experiment that includes some common univariate data generation mechanisms that may be encountered in practice. The rationale is to assess the feasibility of this approach as a possible impetus for extensions to the multivariate case, and to gauge its generalizability potential for creating imputations under some of the broader classes of families mentioned before. Considering that imputation under non-normal densities is a recently emerging notion, which has potential in many research areas, it is important to evaluate its performance in terms of commonly accepted bias and precision measures.

The organization of the rest of the paper is as follows. In section 2, we present a simple algorithm to perform MI under beta and Weibull distributions along with the relevant estimation procedures for obtaining the underlying parameters; we describe a limited simulation study in which we examine the relative improvements over Gaussian imputation on incomplete data sets that exhibit different distributional characteristics; and we explore the behavior of efficiency and accuracy

measures to determine the extent to which the procedures work properly. In section 3, a real data application that ideally suits the proposed method is provided. Section 4 includes concluding remarks, discussion and future directions.

2 Imputation under more flexible distributions

For the purposes of this paper, we focus primarily on incomplete univariate data. Extensions to the multivariate case are discussed in section 4. A key step in the imputation process, which is described subsequently in section 2.2, is estimation of model parameters. Therefore, we begin by briefly going over this background material in the next subsection.

2.1 Estimating the parameters of beta and Weibull densities

Beta and Weibull are two well-known univariate distributions that, depending on the choice of parameter values, can take a variety of distributional forms. This makes them suitable underlying parametric candidates in the imputation process. The major estimation procedures for beta and Weibull densities are as follows.

The probability density function of the beta distribution is

$$f(x|\alpha,\beta)=\frac{(x-a)^{\alpha-1}(b-x)^{\beta-1}}{B(\alpha,\beta)(b-a)^{\alpha+\beta-1}} \quad \text{for} \quad a<x<b; \qquad \min(\alpha,\beta)>0,$$

where α and β are the shape parameters, a and b are the lower and upper bounds, respectively, of the distribution, and $B(\alpha,\beta)$ is the beta function,

$$B(\alpha,\beta)=\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\mathrm{d}t.$$

Given a dataset, the parameters α and β can be estimated by the method of maximum likelihood (ML) and method of moments (MM). Assuming that a and b are known, the ML estimates can be obtained by solving the following set of equations.

$$\psi(\hat{\alpha})-\psi(\hat{\alpha}+\hat{\beta})=\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{x_i-a}{b-a}\right) \quad \text{and} \quad \psi(\hat{\beta})-\psi(\hat{\alpha}+\hat{\beta})=\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{b-x_i}{b-a}\right),$$

where ψ is the digamma function, $\psi(x)=\Gamma'(x)/\Gamma(x)$, and the x_i values are elements of the data vector of length n. The MM estimates are obtained as

$$\hat{\alpha}=\frac{\bar{x}-a}{b-a}\left(\frac{(\bar{x}-a)(b-\bar{x})}{s^2}-1\right), \quad \hat{\beta}=\frac{b-\bar{x}}{b-a}\left(\frac{(\bar{x}-a)(b-\bar{x})}{s^2}-1\right),$$

where $\bar{x}$ is the sample mean and s^2 is the sample variance. The choice between MM and ML estimation can be made based on the relative magnitude of mean squared error (MSE), which is given by

$$MSE=\sum_{i=1}^{n}[\hat{F}(x_i)-F(x_i)]^2,$$

where $F(x_i)$ is the empirical cumulative distribution function, and

$$\hat{F}(x_i) = B^{-1}(\hat{\alpha}, \hat{\beta}) \int_0^{x_i} t^{\hat{\alpha}-1}(1-t)^{\hat{\beta}-1}\mathrm{dt} \quad \text{for} \qquad 0 < x_i < 1$$

(assuming that $a = 0$, $b = 1$).

The density function of Weibull is

$$f(x|\gamma, \delta) = \frac{\delta}{\gamma^{\delta}} x^{\delta-1} \exp\left(-\left(\frac{x}{\gamma}\right)^{\delta}\right), \quad \text{for} \qquad x > 0,$$

where $\gamma > 0$ and $\delta > 0$ are the scale and shape parameters, respectively. ML estimation proceeds by taking the derivatives of the log-likelihood function with respect to γ and δ, which yields

$$\frac{\sum_{i=1}^{n} x_i^{\delta} \log(x_i)}{\sum_{i=1}^{n} x_i^{\delta}} - \frac{1}{\delta} - \frac{1}{n}\sum_{i=1}^{n} \log(x_i) = 0.$$

A standard iterative procedure, such as the Newton–Raphson method, can then be utilized to obtain $\hat{\delta}$. Once this is done, $\hat{\gamma}$ can be found using the equation

$$\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} x_i^{\hat{\delta}}.$$

The MM estimate of δ is based on the coefficient of variation (ratio of the sample standard deviation and mean) which equals $\sqrt{\Gamma(1+\frac{2}{\delta}) - \Gamma^2(1+\frac{1}{\delta})}/\Gamma(1+\frac{1}{\delta})$. Subsequently, the identity $\hat{\gamma} = [\bar{x}/\Gamma(1+\frac{1}{\delta})]^{\delta}$ provides an estimate of γ, where $\bar{x}$ is the mean of the data. Again, one can resort to the MSE to choose between the MM and ML estimates. With an underlying Weibull density, $\hat{F}(x_i) = 1 - \exp(-x_i/\hat{\gamma})^{\hat{\delta}}$.

2.2 Simulation design and imputation algorithm

2.2.1 Complete data generation

Data were generated with six continuous densities (normal, t, lognormal, beta, Weibull and Tukey's gh) that were used in 14 scenarios. The first three are standard distributions, and so their density functions are not included here because of space limitations, and the beta and Weibull densities were mentioned in the previous subsection. Tukey's gh density is not available in closed form and is based on a transformation of standard normal variates (Martinez and Iglewicz, 1984). The transformation is $\mu + \sigma(\exp(gZ) - 1)/g \exp(hZ^2/2)/g$, where μ, σ, g, and h are the location, scale, skewness, and elongation parameters, respectively, and $Z \sim N(0,1)$.

The number of observations, n, in the complete dataset was chosen to be 200, 500, and 1000. The following 14 densities were used in the complete generation process:

1. *Symmetric*: standard normal, standard t with three degrees of freedom, Beta(5,5), and Weibull(1,3.6).
2. *Mode at the boundary*: beta(1,3), Weibull(1,1), and $gh(0,1,1,-0.25)$. (These densities are in fact positively skewed, but 'boundary mode' is a more distinctive and extreme property.)

3. *Positively skewed*: standard lognormal, beta(5,30), Weibull(1,1.5), and $gh(0,1,0.75,0.25)$.
4. *Negatively skewed*: beta(5,1.5), Weibull(1,20), and $gh(0,1,-0.75,0.25)$.

2.2.2 *Missingness mechanism*

Missing values were assumed to be missing completely at random (MCAR). The nonresponse rate was chosen to be 25%, 50%, and 75%.

2.2.3 *Imputation algorithm*

We assume three imputation models for comparison purposes for each incomplete dataset generated. The first one is the normal model, where we create imputations following the standard approach of using a Bayesian predictive model of the missing data given the observed data (Schafer, 1997). For the other two (beta and Weibull models), instead of adopting a Bayesian approach, we account for the parameter uncertainty by obtaining nonparametric bootstrap samples that anchor the subsequent estimation procedure for the parameters of the beta and Weibull distributions, as was done by He and Raghunathan (2006). Denoting the data $Y=(y_1, y_2,\ldots,y_{n_1},y_{n_1+1},\ldots,y_n)^{\mathrm{T}}$, of which the first n_1 elements are observed, and the remaining $n-n_1$ elements are missing, the imputation algorithm is as follows:

1. *Adjust the range*. Transform the observed data, $y_{\mathrm{obs}}=(y_1,y_2,\ldots,y_{n_1})^{\mathrm{T}}$, to lie in (0,1) for beta imputation. (We choose to work with the standard beta distribution, where $a=0$ and $b=1$. This is not a requirement and was done for convenience.) This can be achieved by

$$y^*_{\mathrm{obs}}=\frac{y_{\mathrm{obs}}-\min(y_{\mathrm{obs}})}{\max(y_{\mathrm{obs}})-\min(y_{\mathrm{obs}})}+\epsilon,$$

   where y^*_{obs} denotes the transformed data, and ϵ is a very small positive jitter to ensure that $\min(y_{\mathrm{obs}})$ is positive. For Weibull imputation, the transformation used is a location shift to the right by an amount of $|\min(y_{\mathrm{obs}})|+\epsilon$ if the minimum is negative.
2. Draw a nonparametric bootstrap sample of size n_1 from y^*_{obs}.
3. Estimate the model parameters ((α,β) and (γ,δ) for beta and Weibull distributions, respectively) by the MM and ML methods that were described in section 2.1. Choose the one that yields a smaller *MSE*.
4. Simulate independent variates from these distributions for every missing data point.
5. Back-transform the filled-in data and the transformed observed data to the original scale.
6. Repeat steps 2–5 independently $m=10$ times.

2.2.4 Parameters of interest

We compared the relative performances of normal, beta, and Weibull imputations on six parameters: the mean and five quantiles (5th, 25th, 50th, 75th, and 95th) that are known to be sensitive to model misspecification.

2.2.5 Evaluation criteria

The simulation experiment was repeated $N = 5000$ times for each of the $14 \times 3 \times 3 = 126$ scenarios (combinations of complete data distributions, dataset sizes, and missingness rates, respectively). Evaluation is conducted based on three quantities: (a) Standardized bias (SB) is the relative magnitude of the raw bias to the overall uncertainty in the system. If the parameter of interest is θ, the standardized bias is $100 \times E((\hat{\theta}) - \theta)/\text{SE}(\hat{\theta})$, where SE stands for standard error. If the standardized bias exceeds 50% in a positive or negative direction, then the bias begins to have a noticeable adverse impact on efficiency, coverage and error rates (Demirtas, 2004). (b) Coverage rate (CR) is the percentage of times that the true parameter value is covered in the confidence interval. If a procedure is working well, the actual coverage should be close to the nominal rate (i.e. type I error rates are properly controlled). We regard the performance of the interval procedure to be poor if its coverage drops below 90% (Collins, Schafer and Kam, 2001). (c) Root-mean-square error (RMSE) is an integrated measure of bias and variance. It is considered to be arguably the best criterion for evaluating $\hat{\theta}$ in terms of combined accuracy and precision. $\text{RMSE}(\hat{\theta})$ is defined as $\sqrt{E_\theta[(\hat{\theta} - \theta)^2]}$. Under this specification, SB is a pure accuracy measure, and CR and RMSE are the hybrid measures of accuracy and precision. (For more detailed discussion on this evaluation system, see Demirtas, 2007.)

2.3 Results

Inferences about the mean exhibited little or no discernible differences, hence were not reported. As mentioned, $14 \times 3 \times 3 = 126$ scenarios were considered for the five quantile parameters, leading to $126 \times 5 = 630$ quantities. In Tables 1–3, we report an aggregated version of the results since it is practically impossible to do it separately for each quantity.

Table 1 presents results comparing the normal and beta imputation models. The first column (Distribution) represents the distributional nature of the complete data

Table 1. Comparison of the normal (N) and beta (B) imputation models.

	$\text{CR}_\text{N} < \text{CR}_\text{B}$	$\lvert\text{SB}\rvert_\text{N} > \lvert\text{SB}\rvert_\text{B}$	$\text{RMSE}_\text{N} > \text{RMSE}_\text{B}$	
Distribution	No–Yes	No–Yes	No–Yes	Total
Symmetric	82–98	106–74	125–55	180
Mode at the boundary	17–118	7–128	21–114	135
Positively skewed	36–144	44–136	52–128	180
Negatively skewed	19–116	37–98	33–102	135
Overall	154–476	194–436	231–399	630

Table 2. Comparison of the normal (N) and Weibull (W) imputation models.

	$CR_N < CR_W$	$\|SB\|_N > \|SB\|_W$	$RMSE_N > RMSE_W$	
Distribution	No–Yes	No–Yes	No–Yes	Total
Symmetric	71–109	112–68	136–44	180
Mode at the boundary	26–109	28–107	16–119	135
Positively skewed	24–156	42–138	43–137	180
Negatively skewed	17–118	52–83	19–116	135
Overall	138–492	234–396	214–416	630

Table 3. Independent performances of the three imputation models in terms of coverage rate and absolute standardized bias with acceptable ranges ≥ 0.90 and $< 50\%$, respectively.

Shape	MI model	CR > 0.90	$\|SB\| < 50\%$	Total
Symmetric	Normal	143	140	180
	Beta	161	154	
	Weibull	153	144	
Boundary mode	Normal	35	19	135
	Beta	133	109	
	Weibull	118	72	
Positively skewed	Normal	64	28	180
	Beta	128	91	
	Weibull	125	98	
Negatively skewed	Normal	46	18	135
	Beta	101	82	
	Weibull	92	75	
Overall	Normal	288	205	630
	Beta	523	436	
	Weibull	488	389	

(symmetric, boundary mode, positively and negatively skewed). Aggregated results are presented for each of these distributional forms. Specifically, the second column ($CR_N < CR_B$) indicates the comparative number of cases where the coverage rate (CR) under the normal and beta model are greater than each other. The third column ($|SB|_N > |SB|_B$) denotes a comparison of standardized biases. The fourth column ($RMSE_N > RMSE_B$) concerns which model has smaller RMSEs, and the last column is about totals. To clarify what is presented in the table, let us concentrate on the first row, where we are considering $4 \times 3 \times 3 \times 5 = 180$ quantities (entry in the last column) that arise from the four symmetric densities, three levels of data length, three levels of nonresponse rate and five different quantiles. In the second column, the rate of coverage under the beta model is smaller than under the normal model for 82 quantities out of 180 (the reverse holds for $180 - 82 = 98$ quantities). In the third column a similar comparison is made for the absolute standardized biases, with 106 cases yielding larger biases for the beta model compared with the normal model (conversely, in 74 cases the beta is superior). The fourth column presents results about the magnitude of RMSEs. In summary, we present the results so that the first and second number favor the normal and beta imputation model, respectively. Other rows (e.g. mode at the boundary) represent the same comparisons for the remaining distributional forms. Table 2 is structurally the same as Table 1, with set of entries that compares the normal and Weibull imputation.

Results shown in Table 1 suggest that imputing under a beta model is associated with substantial improvements over Gaussian imputation in terms of coverage rate and absolute standardized bias. Moreover, the beta model yields smaller RMSEs when the mode is at the boundary or the shape is skewed in either direction. The only visible advantage of the Gaussian model appears to be with symmetric densities, although the coverage rate still favors the beta model. Similar conclusions can be drawn from Table 2, where the Weibull model replaces the beta model. One pitfall of this mode of reporting is that the coverage rate or absolute standardized bias can be worse using one model than the other, but the 'inferior' quantity may still be in the acceptable range (≥ 0.90 for CR, $< 50\%$ for |SB|); or a 'superior' quantity may go beyond the acceptable limits. To address this limitation, Table 3 tabulates the *independent* performances of the three imputation models in terms of coverage rate and absolute standardized bias, with acceptable ranges ≥ 0.90 and $< 50\%$, respectively, for each distributional shape. These more elaborate results seem to strengthen the previous finding that imputing under more flexible distributions may significantly enhance the quality of MI inferences for the majority of underlying incomplete data shapes. It is only in the symmetric case that the performance measures appear to be relatively compatible with no clear pattern.

A concern raised by a referee is that no upper bound for the acceptable coverage rate is specified. Coverage rates that are substantially larger than the nominal value (95% in our case) essentially translate to efficiency losses. The upper bound of confidence intervals in binomial proportions depends on the number of simulated replicates (5000), and is 95.6%. However, acceptable upper limit of the coverage rate and undetectable differences in the coverage rates due to the finite replication size are two different concepts; and we are not aware of any published guidelines as to what the upper limit should be in similar simulated missing-data settings. Moreover, an incorrect account of possible precision losses is unlikely because RMSEs are also ascertained and reported in all three tables.

These results are promising in the sense that when the assumption of normality is violated in some fashion, more flexible distributions deliver better performance in most of the simulated scenarios considered. Previous studies reported that imputation under the assumption of normality may be a reasonable approach (Demirtas, Freels and Yucel, 2007b) for estimating the mean parameter (we also found congenial results here in this work). However, examining the parameters that are known to be sensitive to model misspecification (quantiles) revealed that when departures from normality are severe, the beta and Weibull imputations appear to yield far superior results, leading to comparable performance with Gaussian imputation when the data follow a symmetric distribution.

3 Real data application

Our real data example comes from behavioral research. A student–parent questionnaire that is designed to explore the relationship between smoking and drinking

behavior among adolescents and factors such as depression scale, parental message and monitoring. The dataset we have access to, consists of six variables: daily cigarette consumption (DAILYCIG), alcohol problem scale (ALCPROB), self-report depression scale for adolescents and parents SCESD and PCESD), parental message about smoking (PMESSAGE), and parental monitoring (PMONIT). DAILYCIG is the average daily smoking rate in the past 7 days. ALCPROB is the scaled average of five items that are pertinent to the time of last drink, quantity and amount of drinks, whether or not getting drunk and getting into trouble. SCESD and PCESD stand for students' (S) and parents' (P) total scores on 20 items in Center for Epidemiological Studies Depression Scale (CESD). The response metric has four categories, 'rarely', 'some of the time', 'occasionally', and 'most of the time' (Radloff, 1977). PMESSAGE is the average of seven items about the frequency of parental smoking messages such as 'smoking gives you cancer', 'smoking is addictive', etc. The response metric includes 'never','once or twice', and 'several times' (Kodl and Mermelstein, 2004). PMONIT is the degree of involved/vigilant parental monitoring as measured by the average of parents' three responses that are given to questions about knowing what their child does after school, if (s)/he does something wrong, and who (s)/he is with when (s)/he is away from home, where the response metric has four categories ranging from 'never' to 'always' (Ge *et al.*, 2004).

Of 1264 adoscelents in the study, about 20% of people do not have observations in PCESD, PMESSAGE, and PMONIT; the other three variables are mostly observed. A comparison of the mean profiles among participants that have complete set of values and the ones whose responses are missing for the three variables mentioned above shows that average profiles for nearly completely observed variables are very similar across the two groups, suggesting that MCAR assumption may be reasonable. The number of observed values per each variable and some descriptive statistics are given in Table 4; and histograms of PCESD, PMESSAGE, and PMONIT are shown in Figure 1. All three variables have non-normal features: the modes of PCESD and PMESSAGE are at the boundary, and PMONIT has skip patterns given the nature of data collection. For these variables, we calculated 5th, 25th, 50th, 75th, and 95th quantiles based on the original data. Subsequently, we applied the three imputation models that were described in section 2.2 to the $m=10$ bootstrap samples, separately for each variable under consideration, with an assumption that the missingness mechanism is MCAR. Under MCAR, observed data and imputed values should have similar distributional properties. Averaged results for the

Table 4. Descriptive statistics on six variables in the system.

Variable	n_1	Mean	Std. Dev.	Minimum	Maximum
DAILYCIG	1233	0.4808921	1.5519367	0	18.57
ALCPROB	1263	3.6223515	1.6739107	1	7.30
SCESD	1260	16.8263810	9.7974900	0	53.00
PCESD	1018	10.8210118	8.9243597	0	53.00
PMESSAGE	1022	2.6377593	0.5028218	1	3.00
PMONIT	1020	3.2204314	0.5372196	1	4.00

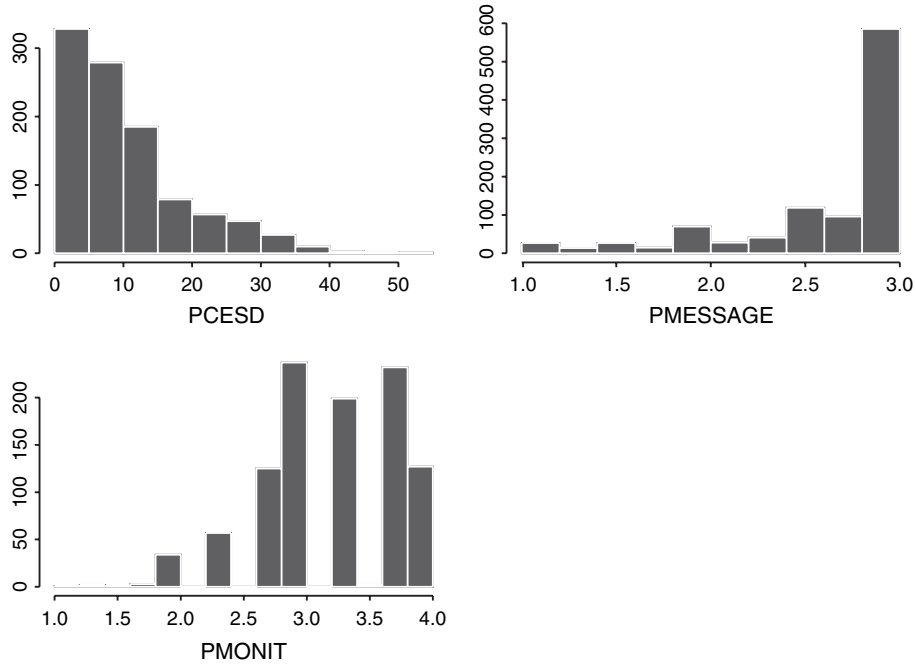

Fig. 1. Histogram of the three incomplete variables.

Table 5. Comparison of the quantiles of observed data and imputed portion across 10 imputations under the normal, beta, and Weibull models.

Variable	Quantile	Original data	Normal	Beta	Weibull
PCESD	5th	0	−3.701293	0.694861	0.871124
	25th	4	4.787408	3.707578	3.778772
	50th	9	10.813007	8.576269	8.212498
	75th	15	16.830011	15.873737	15.088093
	95th	29	25.449233	28.428649	29.591730
PMESSAGE	5th	1.43	1.819485	1.486685	1.748191
	25th	2.43	2.300578	2.451136	2.310018
	50th	2.86	2.640129	2.894021	2.673025
	75th	3.00	2.980188	2.993039	2.997765
	95th	3.00	3.458390	2.999980	3.399562
PMONIT	5th	2.33	2.346608	2.199817	2.267987
	25th	3.00	2.862594	2.884083	2.882017
	50th	3.33	3.221125	3.323434	3.267762
	75th	3.67	3.582343	3.651867	3.606597
	95th	4.00	4.095061	3.905260	4.021580

incomplete portions of the data across 10 imputed datasets for five quantiles are tabulated in Table 5. For the total of 15 quantities, ranking the three imputation models, with 1 is the best and 3 is the worst in terms of closeness to the observed data quantiles, the average ranks are 2.866667, 1.4, 1.733333 for normal, beta, and Weibull models, respectively. Our real data example results seem to strongly support the findings in section 2.

4 Discussion

There are a few limitations that need to be addressed. First, while we recognize that real incomplete data often include many variables, our focus was on univariate data. We view this as a potential building block for more realistic situations. The behavior of the third and fourth moments typically requires more modeling flexibility in terms of the area covered in the symmetry-elongation plane as well as the association among variables. This work serves as an initial feasibility study for assessing the generalizability potential to the multivariate settings. On a related note, although the beta and Weibull distributions are capable of picking some data trends that are unlikely to be captured by a normal model, they do not cover the entire symmetry-elongation plane. Nevertheless, considering the relative gains presented in this paper, it provides an indication that generalized families of densities can lead to further improvements. Second, the assumed missingness mechanism (MCAR) is generally too simplistic for real-life applications. However, our purpose was not to conduct a sensitivity analysis with respect to the mechanism that leads to the observed data. Rather, the current paper was motivated by how tenably MI inferences can be conducted with assumed non-Gaussian continuous distributions. Again, all new research has to start from some point. Third, one can question the way we report the results on the grounds that it is overly aggregated. Although this argument has some validity, restricted space does not permit more detailed reporting and the tabulated results are adequate for conveying the primary message of this paper. Fourth, as correctly pointed out by a reviewer, parameter uncertainty was taken into account through nonparametric bootstrap samples. For univariate data, sampling with replacement is equivalent to using the inverse CDF (cumulative distribution function) method on the empirical CDF. Although it is a reasonable approach for moderate and large sample sizes, it may lead to unacceptable random samples in the small-sample case where the number of distinct data values is limited, which in turn may generate an unduly large degree of variability among parameter estimates in the subsequent step. Demirtas and Hedeker (2006) argue that in order to circumvent this potential complication, two simple ideas from nonparametric density estimation can be employed based on using a smoothed variant of the empirical CDF: (1) Binning the data and forming a frequency polygon, and using the inverse CDF approach on the resulting distribution function which is a piecewise quadratic polynomial; (2) Connecting the jump points of the empirical CDF with line segments to form a piecewise linear function. Finally, our simulation setup had to be in manageable limits and does not span every imaginable scenario that may arise in practice. However, we believe that it is sufficiently comprehensive to demonstrate the superiority of non-Gaussian imputation models in most cases.

The assumption of multivariate normality along with the Bayesian paradigm has often been regarded as a statistically defensible way of creating multiply imputed data sets for continuous data. While it is a convenient assumption and it has been shown to work well in some settings (e.g. with a large number of subjects), it is constructive to move the practice of MI to other distributions that cover a broader

range of the third and fourth moments. In an attempt to go beyond the realm of normality to adequately model distributional properties that are not accommodated by a Gaussian model, there has been a growing interest in non-normal distributions. This work was motivated by the premise that the MI framework may be amenable to more general families that were mentioned in section 1. Forming Bayesian predictive distributions under these families may be a formidable task in a complex multivariate setting, however, it appears to be a potentially fruitful future research area. Although our work is limited to incomplete univariate data and MI under two relatively simple densities, we believe that it has prospects to be extended to the multivariate case as well as to more sophisticated and flexible distributions.

Acknowledgements

We thank Dr Robin Mermelstein for access to the data; part of this work was supported by National Cancer Institute grant 5PO1 CA98262.

References

Belin, T. R., M. Y. Hu, A. S. Young and O. Grusky (2000), Using multiple imputation to incorporate cases with missing items in a mental health services study, *Health Services and Outcome Research Methodology* **1**, 7–22.

Burr, I. W. (1942), Cumulative frequency functions, *Annals of Mathematical Statistics* **13**, 215–232.

Collins, L. M., J. L. Schafer and C. H. Kam (2001), A comparison of inclusive and restrictive strategies in modern missing data procedures, *Psychological Methods* **6**, 330–351.

Demirtas, H. (2004), Simulation-driven inferences for multiply imputed longitudinal datasets, *Statistica Neerlandica* **58**, 466–482.

Demirtas, H. (2005a), Bayesian analysis of hierarchical pattern-mixture models for clinical trials data with attrition and comparisons to commonly used ad-hoc and model-based approaches. *Journal of Biopharmaceutical Statistics* **25**, 383–402.

Demirtas, H. (2005b), Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out, *Statistics in Medicine* **24**, 2345–2363.

Demirtas, H. (2007), Practical advice on how to impute continuous data when the ultimate interest centers on dichotomized outcomes through pre-specified thresholds, *Communications in Statistics – Simulation and Computation* **36**, 871–889.

Demirtas, H. and D. Hedeker (2006), Comment on "Tukey's gh distribution for multiple imputation", *American Statistician* **60**, 348–349.

Demirtas, H. and D. Hedeker (2007), Gaussianization-based quasi-imputation and expansion strategies for incomplete correlated binary responses, *Statistics in Medicine* **26**, 782–799.

Demirtas, H. and J. L. Schafer (2003), On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out, *Statistics in Medicine* **22**, 2553–2575.

Demirtas, H., L. M. Arguelles, H. Chung and D. Hedeker (2007a), On the performance of bias-reduction techniques for variance estimation in approximate Bayesian bootstrap imputation, *Computational Statistics and Data Analysis* **51**, 4064–4068.

Demirtas, H., S. A. Freels and R. M. Yucel (2007b), Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment, *Journal of Statistical Computation and Simulation* **77**, 000–000 (in press).

Fleishman, A. I. (1978), A method for simulating non-normal distributions, *Psychometrika* **43**, 521–532.

Ge, X., R. D. Conger, F. O. Lorenz and R. L. Simons (1994), Parent's successful life events and adolescent depressed mood, *Journal of Health and Social Behavior* **35**, 28–44.

Genton, M. G. (Ed.) (2004), *Skew-elliptical distributions and their applications: a journey beyond normality*, Chapman and Hall/CRC, Boca Raton, FL.

He, Y. and T. E. Raghunathan (2006), Tukey's gh distribution for multiple imputation, *The American Statistician* **60**, 251–256.

Horton, J. H. and S. R. Lipsitz (2001), Multiple imputation in practice: Comparison of software packages for regression models with missing variables, *The American Statistician* **55**, 244–254.

Johnson, N. L. (1949), Systems of frequency curves generated by methods of translation, *Biometrika* **36**, 149–176.

Kodl, M. M. and R. Mermelstein (2004), Beyond modeling: Parental practices, parental smoking history, and adolescent cigarette smoking, *Addictive Behaviors* **29**, 17–32.

Little, R. J. A. and D. B. Rubin (2002), *Statistical analysis with missing data*, 2nd edn, Wiley, New York.

Liu, C. (1995), Missing data imputation using the multivariate t distribution, *Journal of Multivariate Analysis* **53**, 139–158.

Martinez, J. and B. Iglewicz (1984), Some properties of the Tukey g and h family of distributions, *Communications in Statistics – Theory and Methods* **13**, 359–369.

Morgenthaler, S. and J. W. Tukey (2000), Fitting quantiles: doubling, HR, HQ, and HHH distributions, *Journal of Computational and Graphical Statistics* **9**, 180–195.

Parrish, R. S. (1990), Generating random deviates from multivariate Pearson distributions, *Computational Statistics and Data Analysis* **9**, 283–295.

Radloff, L. S. (1977), The CES-D scale: a self-report depression scale for research in the general population, *Applied Psychological Measurement* **1**, 385–401.

Rubin, D. B. (1996), Multiple imputation after 18+ years (with discussion), *Journal of the American Statistical Association* **91**, 473–520.

Rubin, D. B. (2004), *Multiple imputation for nonresponse in surveys*, Wiley Classic Library, New York.

Schafer, J. L. (1997), *Analysis of incomplete multivariate data*, Chapman & Hall, London.

Schafer, J. L. (1999), Multiple imputation: a primer, *Statistical Methods in Medical Research* **8**, 3–15.

Schimert, J., J. L. Schafer, T. Hesterberg, C. Fraley and D. B. Clarkson (2001), *Analyzing data with missing values in S-plus*, Data Analysis Products Division, Insightful Corp., Seattle, WA.

Van Buuren, S. and C. G. M. Oudshoom (2000), *MICE V1.0 Users Guide*, Leiden: TNO Preventie en Gezonheid, TNO/PG/V GZ 00.038.

Van Buuren, S., H. C. Boshuizen and D. L. Knook (1999), Multiple imputation of missing blood pressure covariates in survival analysis, *Statistics in Medicine* **18**, 681–694.

Received: October 2006. Revised: June 2007.