APPLICATION OF GENERALIZED ITERATIVE BAYESIAN SIMULATION METHODS TO ESTIMATION AND INFERENCE FOR COARSENED HOUSEHOLD INCOME AND ASSET DATA

Steven G. Heeringa, University of Michigan
Institute for Social Research, University of Michigan, P.O. Box 1248, Ann Arbor, MI 48106-1248

1. Introduction

Survey questions that ask respondents to report amounts -- particularly dollar values for financial variables such as income and assets, liabilities, transfers -- are subject to high rates of item missing data. (Juster and Smith, 1995). As an alternative to simply accepting high rates of item missing data for financial variables, researchers are making increased use of special questionnaire formats that are designed to collect an interval-scale observation whenever a respondent is unable or unwilling to provide an exact response to a financial amount question (Heeringa, 1993). Loosely termed "bracketing questions," these new question formats are a type of the more general class of unfolding question sequences that are developed for improving survey measurements of complex characteristics.

The use of interval scale measures for financial items is not new to survey research. The simple income questions included in many survey questionnaires are often designed to measure amounts on an interval scale (e.g., $0-4999, $5000-$14,999, etc.). In face-to-face interview situations, "show cards" or other visual devices enable respondents to map an underlying cardinal-valued response item onto an interval or ordinal scale. In surveys where cardinal-scale measurement of financial variables is necessary or preferred, the University of Michigan Survey Research Center (SRC) has historically provided its interviewers with a "range card" which enabled them to record an interval scale response code for financial amount items. Unlike show cards, the range card was not designed to be used each time the question was asked but served as an interviewer aid in cases where it was clear that the respondent would not report an actual amount. To avoid confusion on the part of the interviewer, a single set of fixed range card categories was applied to all financial measures regardless of their underlying distribution in the population. In large part, the frequency and accuracy of range card responses to financial amount items was determined by the individual interviewer.

Bracketing question sequences for measuring financial variables first appeared in the special wealth supplement to the 1984 Panel Study of Income Dynamics (PSID, see Curtin, Morgan and Juster, 1989). Bracketed measurement of 1984 PSID households' financial assets served to: 1) standardize the process for recovering interval scale observations for missing amounts; 2) adapt the interval scales to the population distribution for the financial variable of interest; and 3) enable the collection of interval scale measures in a telephone interview format. The use of bracketing question sequences was repeated in the 1989 and 1993 wealth supplements to the PSID. This paper will draw heavily on data and field experience with bracketed question items used in the more recent 1992 Health and Retirement Survey (HRS). Through the use of special question formats the rate of completely missing data for HRS asset amount variables is significantly reduced; however, the resulting measures are a mixture of single valued responses, "bracketed" or interval valued responses, and completely missing data.

The special question format that is the direct cause of the bracketed data problem has proved to be a very useful tool for addressing the serious missing data problems that are common for income and asset variables. The technique, or some variant of it, is already being used on other major surveys of household financial characteristics. Now comes the question of how to best use these coarsened data in multivariate estimation and inference, or alternatively in imputation of item missing data for public use data sets. For the multivariate problem, Heeringa (1993) initially proposed the use of the general location model (Little and Rubin, 1987) and a generalized iterative Bayesian (GIBS) algorithm to derive estimates of model parameters or to perform multiple imputation for the bracketed response data. The following paper revises and extends that early work to propose a different GIBS approach which corrects many of the short-comings of the method proposed in the 1993 paper. The revised approach incorporates several suggestions offered by Little (1993) in his discussion of the earlier approach.

Including this introduction, the paper is organized in four major sections. Section 2 reviews the bracketed response data problem using as examples data from Wave 1 of the Health and Retirement Survey (HRS). Section 3 outlines a coarsened data model (Heitjan and Rubin, 1991) for these forms of bracketed response data. Section 4 describes a modified version of the GIBS data augmentation method (Tanner and Wong, 1987) that is adapted to this special "coarsened data" problem and illustrates how the algorithm can be used

to simulate the effects of nonignorable coarsening in the response process.

2. The Data: Bracketed Response Question Formats

Figure 1 illustrates the format of the bracketing question sequence for two asset items: equity in a business and combined value of IRA and Keogh accounts. For these and seven other key asset items, if a respondent could not recall or refused to report the exact value for the item, the HRS Wave 1 questionnaire followed up with a short sequence of questions designed to "bracket" the underlying response value. The question sequences open by asking if the household owns the asset (e.g., a business). If the asset is owned, its exact value is requested. If the exact value is not reported, the questionnaire routes the respondent through a series of dichotomous response questions which attempt to bracket the value of the asset. Taking the business asset and IRA/Keogh account value question sequences as examples, the finest level of bracketing attainable through the HRS Wave 1 questions is shown in Table 1 below.

Routing the respondent through the nested series of bracketing questions does not guarantee that a specific bracket will be identified for the unreported amount. In some cases, no additional information will be obtained. In other cases, the responses will indicate that the true value lies in one of three brackets, but not precisely which of the three brackets. By example, a respondent may indicate that the value of his IRA or Keogh account is >= $25,000 but cannot/will not indicate if it is $25,000-$49,999, $50,000-$99,999, or $100,000+.

Table 2 summarizes the HRS Wave 1 data problem for each of the nine household assets. The left-hand panel of Table 2 identifies the individual asset components in question. The central panel, labeled "Does item apply?", provides estimates of the percentage of HRS Wave 1 sample households (unweighted) that reported having each asset (i.e., a nonzero amount value is assumed). For example, of the n=7608 respondent households included in this summary, 23.1% report owning real estate other than their personal residence. For households that report owning a particular asset or having a particular type of debt, the right-hand panel of Table 2 describes the distribution of response types: actual value, bracketed value,[1] range card value, or missing data value.

Among financial assets, the percentage of actual value reports ranges from 67.4% for stocks and mutual funds to 87.4% for combined value of vehicles and other personal property. Depending on the asset, the percentage of bracketed responses ranges from 8.2% for property to 21.3% for business value. Even though a bracketing question sequence was provided for these asset items, from 2.4% to 6.5% of bounded response values were recorded as choices from the range card. The rates of completely missing data -- proportions of cases where no real information on bounding values is available -- range from 1.9% of responses for the vehicle and property question to 10.6% for value of bonds.

The bracketed data problem is potentially compound. One aspect of the coarsening has its origin in deliberate questionnaire design formats that are intended to recover partial information when respondents cannot/will not report the desired amount value. In addition to this expected coarsening of responses, it appears that further undesired coarsening of responses may occur in the respondent reports of actual amounts that exhibit a high degree of "heaping" at selected values (i.e., 1000s, 10,000s, etc.). Only the coarsening due to bracketing questions is considered here.

3. The Coarse Data Model (Heitjan and Rubin, 1991)

The coarsened data model (Heitjan and Rubin, 1991) provides a framework for approaching the problem of estimation for bracketed response question data. The general statistical model for coarse data assumes a continuous underlying variable, x, and a parameter vector, θ, for which estimation and inference are of interest:

$$X \sim f(x|\theta)$$

Examples from statistical practice might be distributions of taxonomic measurements on human heights (Wachter and Trussell, 1982) that are assumed to follow a normal distribution, post-operative survival times assumed to follow an exponential or other gamma distribution (Heitjan, 1993), or an income or asset distribution that follows a mixture of a lognormal and Pareto distributions (Aigner and Goldberger, 1970).

3.A The Coarsened Data Likelihood

In addition to the underlying continuous variable of interest, each observational unit holds a value for a reporting variable, G, which determines the state of coarsening with which the individual X's will be reported. Conditional on the variable, x, and the parameter vector, γ, G is distributed as:

$$G \sim h(g|x,\gamma)$$

$$G: X \to Y$$

As indicated, the random variable G for each subject determines the mapping of the underlying continuous variable, x, to the observed variable y.

Figure 1

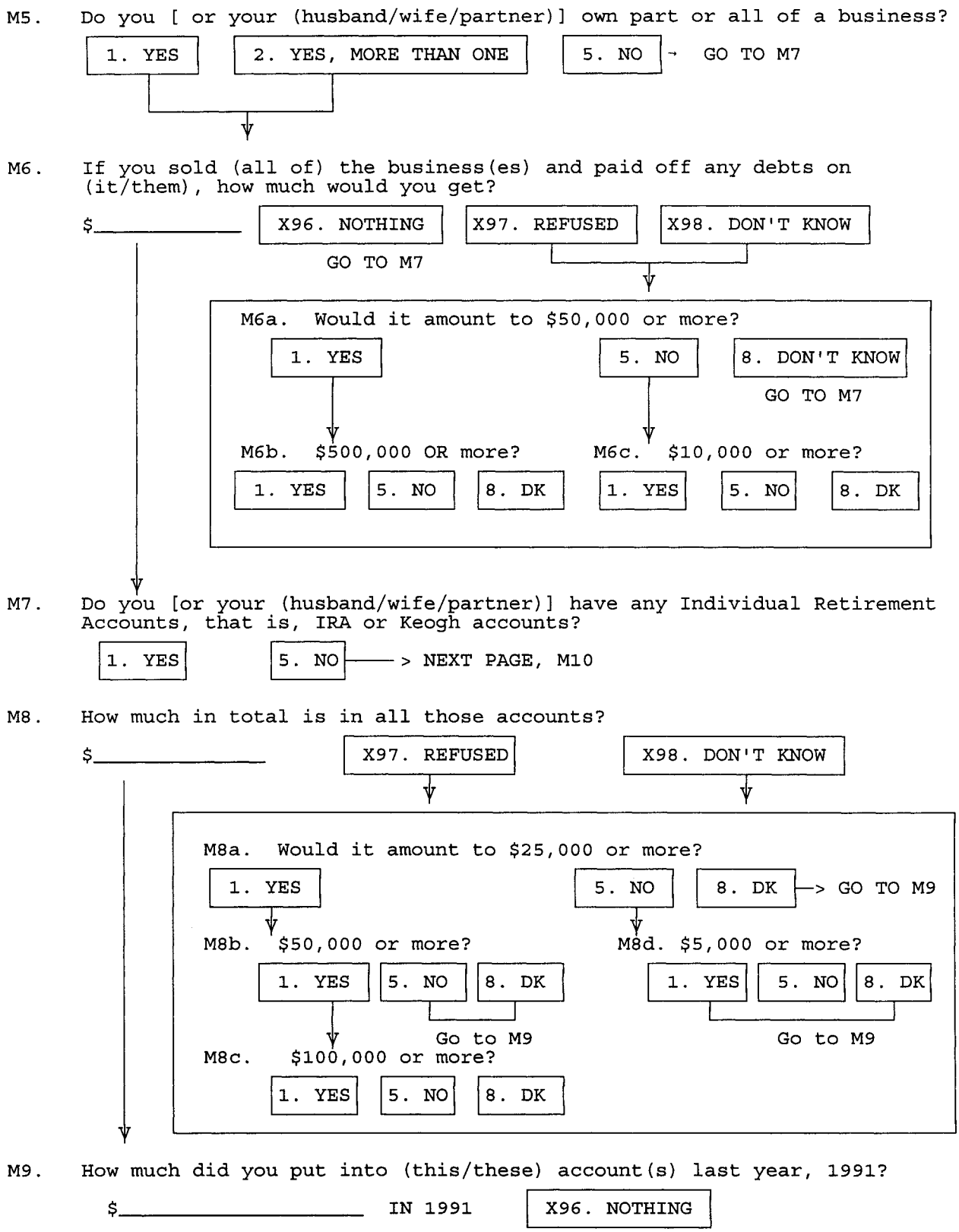

M5. Do you [or your (husband/wife/partner)] own part or all of a business?
1. YES
2. YES, MORE THAN ONE
5. NO → GO TO M7
M6. If you sold (all of) the business(es) and paid off any debts on (it/them), how much would you get?
$_____
X96. NOTHING
GO TO M7
X97. REFUSED
X98. DON'T KNOW
M6a. Would it amount to $50,000 or more?
1. YES
5. NO
8. DON'T KNOW
GO TO M7
M6b. $500,000 OR more?
1. YES
5. NO
8. DK
M6c. $10,000 or more?
1. YES
5. NO
8. DK
M7. Do you [or your (husband/wife/partner)] have any Individual Retirement Accounts, that is, IRA or Keogh accounts?
1. YES
5. NO ——> NEXT PAGE, M10
M8. How much in total is in all those accounts?
$_____
X97. REFUSED
X98. DON'T KNOW
M8a. Would it amount to $25,000 or more?
1. YES
5. NO
8. DK —> GO TO M9
M8b. $50,000 or more?
1. YES
5. NO
8. DK
Go to M9
M8c. $100,000 or more?
1. YES
5. NO
8. DK
M8d. $5,000 or more?
1. YES
5. NO
8. DK
Go to M9
M9. How much did you put into (this/these) account(s) last year, 1991?
$_____ IN 1991
X96. NOTHING

Table 1
Examples of Response Bracket Ranges for HRS Asset Items[2]

Bracket	Business Value Response	IRA, Keogh Response
1	$1 - $9,999	$1-4,999
2	$10,000 - $49,999	$5,000 - $24,999
3	$50,000 - $499,999	$25,000 - $49,999
4	$500,000 +	$50,000 - $99,999
5	Not applicable	$100,000 +

Table 2
HRS Wave 1 Net Worth Components
Distribution of Responses by Response Type (n = 7608 respondent households)

Household Asset	Does Item Apply?				If Item Applies To Household					
					Total		Actual Value	Bracketed Value	Range Card Value	Missing Value
	Total	Yes	No	DK	n	%				
A: Real Estate (not home)	100%	23.1%	76.3%	0.6%	1759	100%	74.7%	16.2%	5.9%	3.4%
A: Vehicles, Pers Prop.	100%	-	-	-	7608	100%	87.4%	8.2%	2.4%	1.9%
A: Business	100%	16.1%	83.4%	0.5%	1226	100%	68.2%	21.3%	4.7%	5.8%
A: IRA, Keogh	100%	37.1%	62.2%	0.7%	2825	100%	73.7%	16.3%	5.2%	4.9%
A: Stock, Mutual Funds	100%	26.5%	72.6%	0.9%	2015	100%	67.4%	21.0%	5.6%	6.0%
A: Checking, Savings	100%	77.5%	21.5%	1.0%	5895	100%	73.3%	16.4%	5.1%	5.2%
A: CDs, Sav Bonds, T-Bills	100%	24.6%	74.3%	1.1%	1870	100%	70.6%	16.2%	6.4%	6.8%
A: Bonds	100%	5.9%	93.2%	1.0%	445	100%	69.7%	13.3%	6.5%	10.6%
A: Other Assets	100%	15.0%	83.9%	1.1%	1143	100%	72.1%	16.3%	5.3%	6.4%

The mapping of x to y is not a 1:1 transformation of the underlying variable. For example, the complex pattern of coarsening observed in the bracketed response data could be summarized by a multinomial coarsening variable with three categories:

$$Y_i = \begin{cases} X_i & G_i = 0 \\ (X_{LB}, X_{UB}] & G_i = 1 \\ \textit{Missing} & G_i = 2 \end{cases}$$

If the realized value of g for subject i is 0, the coarsened variable is equal to the actual amount (the original intent of the question). For $G_i=1$, the amount is not reported but is indicated to lie in a specific interval of the full range of x, $(X_{LB}, X_{UB}]$. Complete missing data results in the case where $G_i=2$.

Conditional on x and g, the distribution of y is degenerate:

$$r(y \mid x,g) = \begin{cases} 1 & y = Y(X,G) \\ 0 & y \neq Y(X,G) \end{cases}$$

The conditional distribution of y given x and the coarsening model parameters, γ, is obtained by integrating g out of the joint distribution of y and g:

$$k(y \mid x,\gamma) = \int_{\Gamma} r(y \mid x,g) \cdot h(g \mid x,\gamma) \cdot dg$$

A complete likelihood function for the estimation of θ that reflects: 1) the coarsened form of x; and 2) the stochastic influence of the reporting or coarsening function $h(g \mid x,\gamma)$ is:

$$L_C(\theta,\gamma \mid y) = \int_{y} f(x \mid \theta) \cdot k(y \mid x,\gamma)\, dx$$

3.B. Coarsened at Random (CAR), Ignorability

Heitjan and Rubin (1991) define the data y to be coarsened at random (CAR) if for each possible value of γ, the conditional density $k(y|x,\gamma)$ is constant for all x that can be mapped into the coarsened variable y. For example, a bracketed response variate y is CAR if the probability of an interval scale response is uniform over the range $I_x = (X_L, X_u]$. If y is CAR, $k(y|x,\gamma)$ can be factored out of the integral expression for $L_C(\theta,\gamma/y)$ and contributes only a product of scaling constants to the likelihood for θ.

Extending Rubin's (1976) result for the special case of the complete missing data problem, Heitjan and Rubin (1991) provide Theorem 1 which states that for distinct parameter sets θ and γ, likelihood ratio tests and Bayesian inference based on L_C are equivalent to those based on the simpler form of the likelihood:

$$L_G(\theta|y) = \int_{\Xi} r(y|x,\theta) \cdot f(x|\theta)\,dx$$
$$= \int_{y} f(x|\theta)dx$$

where: Ξ = the sample space of x

Little and Rubin (1987) describe the use of the E-M algorithm to estimate the parameters of this grouped data likelihood.

3.C Bracketed Response Measures as a Coarse Data Problem

Focusing on a single variable, the observed responses to bracketed response question items can be represented as a coarsening, Y:(X,G), of the underlying variable x. To simplify this discussion, we will focus on a model of coarsening in which the mapping of x-->y results in 1 of 2 states -- uncoarsened and coarsened.

$$Y_i = \begin{cases} X_i & G_i = 0 \text{ uncoarsened} \\ (X_L, X_U] & G_i = 1 \text{ coarsened} \end{cases}$$

Complete missing data will be treated as a special case of coarsening on the range $(0, +\alpha)$.

Following the economists' treatment of limited dependent variables (Maddala, 1983), zero values will be treated as censoring on the interval $(-\infty, 0]$. This approach was suggested by Little (1993). A similar technique was used by Little and Su (1987).

3.D Is Bracketing CAR?

To realize the full value of the added information gleaned by this bracketed response question method, steps must be taken to understand how best to analyze[3] the coarsened data. One important step in this understanding is to learn more about the coarsening mechanism. A major question is whether the bracketing represents a CAR mechanism. Table 3 provides indirect support for the hypothesis that the bracketing process is not CAR. Table 3 also suggests that the completely missing asset data are not missing at random (MAR).

Table 3 focuses on HRS household reports of stock and mutual fund value. In the final survey data set, a total of 2015 HRS Wave 1 sample households reported owning stock or mutual funds. At the initial question, 1339 respondents reported actual amounts, 112 reported a range card interval response, 413 said they did not know (DK) the actual amount, and 131 refused (REF) to report an exact dollar amount. When asked the follow-up bracketing question sequence, 87.9% (363 of 413) of the original DK responses provided interval scale responses for the amount. The bracketed follow-up questions also elicited an interval scale response from 59 of 131 (45.0%) of the original refusers. While there are no guarantees concerning the sample properties of the subsets of initial DKs and REFs that provided bracketing information, Table 3 shows a very clear pattern. The observed distribution of the initial DKs to the 5 amount brackets is very similar to the grouped distribution for households who reported actual amounts. In comparison to the actual value reporters and the DKs, initial refusals appear to distribute more heavily to the middle and upper brackets.

Furthermore, initial refusals are much more likely to provide no information at all. These data suggest that the coarsening itself may reflect a mixture of two processes, one originating in the lack of information (DKs) and the second due to other factors (REFs) that are clearly associated with the value of the asset itself. Both the propensity to provide a bracketed response and the probability of completely missing data appear to increase with the value of the asset.

4. Data Augmentation for Multivariate Data With Nonignorable Coarsening

Heitjan and Rubin (1991) have outlined a theoretical framework for estimation and inference based on coarsened data. Section 3 attempted to place a real data problem encountered in bracketed response survey questions into that theoretical framework. If the data are CAR or if the model for the coarsening mechanism can be identified, models involving these coarsened variables can be studied using maximum likelihood or generalized iterative Bayesian GIBS) approaches. To fit general multivariate models to the coarsened data described in Section 2, a ML or GIBS algorithm must be able to handle three features of the coarsened data:

i) zero values in the multivariate vector;

ii) the interval censoring of individual variable values; and

iii) nonignorable coarsening.

Table 3
HRS Wave 1 Stock and Mutual Funds Amount
Distribution of Responses to Brackets*

	Stocks and Mutual Funds						
Response Bracket	**Amount Range**	**Reported Actual Amount**		**Don't Know Actual Amount**		**Refused Actual Amount**	
		n	**%**	**n**	**%**	**n**	**%**
1	$1 - $4999	298	22.3%	84	23.2%	6	10.2%
2	$5K - $24.9K	455	34.0%	123	33.9%	12	20.3%
3	$25K - $99.9K	362	27.0%	109	30.0%	15	25.4%
4	$100K - $500K	199	14.8%	40	11.0%	19	32.3%
5	$500K +	25	1.9%	7	1.9%	7	11.9%
With Bracket		1339	100%	363	100%	59	100%
Without Bracket		0		50		72	
Total Cases		1339		413		131	

*Table does not reflect n=112 respondents who were permitted to use a range card response to the initial query.

This section proposes an approach based on the data augmentation method of Tanner and Wong (1987) which simplifies the estimation problem under conditions (i)-(iii) above.

The recent literature on iterative simulation methods describes at least two approaches that could be considered for the bracketed response data problem. The two approaches are: 1) the data augmentation algorithm (Tanner and Wong, 1987) and 2) Monte Carlo implementation of the EM algorithm (Wei and Tanner, 1990). The two approaches are similar in that each uses an iterative, "imputation-based" algorithm to evaluate the posterior distribution of the model parameters, θ. Under the probability model for the complete data and a suitable choice of a prior for θ, data augmentation calculates the full posterior distribution. The Monte Carlo EM approach is designed to solve for the mode of the posterior, although Wei and Tanner (1987) describe a "poor man's data augmentation" extension of the Monte Carlo EM algorithm which can be used to estimate the shape of the complete posterior distribution.

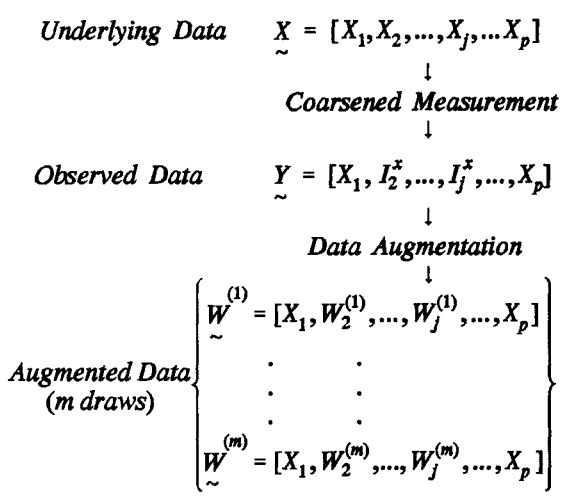
4.A Data Augmentation for the Bracketed Response Data

Adopting a Bayesian approach to the estimation and inference problem, the objective is to compute the posterior distribution, $p(\theta|y)$, where y is the observed, coarsened data. The complexity of the distribution for the observed data rules out a simple, direct evaluation of this posterior using conventional Bayes' methods. However, as shown in Tanner and Wong (1987), it is possible to augment the observed data with additional data that allows the problem to be solved by an iterative algorithm.

For the bracketed response data problem considered in this paper, the required augmentation is a subvector of regression imputations, $\underset{\sim}{W}$, with one element for each element of $\underset{\sim}{X}$ that is coarsened in the measurement process. The regression imputations are constrained to reflect the information on the potential range of x, $(X_L, X_U]$, provided by the observed data vector, $\underset{\sim}{Y}$.

Figure 2 is a schematic illustration of the structure and relationship of the underlying, observed and augmented data records.

Figure 2
Example Data Structures



To explain the application of the data augmentation algorithm to the bracketed response data problem, we need the following expression for the posterior density of θ:

$$p(\theta \mid y) = \int_W p(\theta \mid w,y)p(w \mid y)\,dw$$

where:

$$p(w \mid y) = \int_\theta p(w \mid \phi,y)p(\phi \mid y)\,d\phi$$

is the posterior predictive density for the augmented data, w, given the observed data, y, and a set of parameters, ϕ in Θ which define the regression of w on y.

Substituting this expression for p(w|y) and changing the order of integration, Tanner and Wong provide the following integral equation for p(θ|y):

$$p(\theta \mid y) = \int_\theta K(\theta,\phi)g(\phi)\,d\phi,$$

$$\text{where } K(\theta,\phi) = \int_W p(\theta \mid w,y)p(w \mid \phi,y)\,dw$$

The iterative data augmentation algorithm outlined by Tanner and Wong applies the method of successive substitution to evaluate this integral equation for the desired posterior p(θ|y). At each iteration, the integral K(θ,ϕ) is evaluated by Monte Carlo methods using multiple draws (imputations) from the posterior predictive distribution, p(w|ϕ,y).

The proposed data augmentation algorithm has three basic steps:

DA Step 1: Generate a sample value of ϕ using the current iteration's version of the posterior density for θ, $p^{(i)}(\theta|y)$. Assume that the distribution of the underlying data, x, is multivariate log-normal, $f(\ln(x);\theta) \sim N(u,\Sigma)$. Given this multivariate normal distribution for the natural log transformation of the underlying income or asset measures, the conjugate prior for $\theta = (\mu,\Sigma)$ is a normal-inverse Wishart distribution (Schaffer, 1991). The diffuse, multivariate Jeffreys prior for $\theta = (\mu, \Sigma)$ is:

$$\pi(\theta) \propto |\Sigma|^{-\left(\frac{p+1}{2}\right)}$$

When the complete data likelihood for $\ln(x)$ is multivariate normal, both the conjugate and Jeffreys prior yield a posterior, p(θ|y), that is also normal-inverse Wishart. For example, the complete data posterior for the improper Jeffreys prior is:

$$p(u \mid \Sigma,y) = N(\bar{y},n^{-1}\Sigma)$$

$$p(\Sigma^{-1} \mid y) = W(n-1,(nS)^{-1})$$

The generation of samples (required in Step 1 of the DA algorithm) from the normal-inverse Wishart posterior is straightforward (Schaffer, 1991).

DA Step 2: Generate a sample of size m of the desired augmented data from the predictive posterior density p(w|y), $\{w^{(1)}, w^{(2)},\ldots,w^{(m)}\}$. The imputation step of the DA algorithm requires random draws from the posterior predictive density:

$$p(w \mid y) = \int_\theta p(w \mid \phi,y)p(\phi \mid y)\,d\phi$$

Given the current draw of θ from the normal-inverse Wishart density, $p^{(i)}(\theta|y)$, the predictive posterior distribution $p(w|\theta^{(i)},y)$ is a multivariate normal linear regression of w on y. (The regression parameters, ϕ, in the integral equation are a reparameterization of θ, the multivariate mean vector and variance/covariance matrix). Schaffer(1991) outlines an algorithm based on the sweep operator which efficiently computes the parameters of this predictive distribution for each pattern of missingness (i.e., coarsening) in the multivariate data vector. The algorithm also includes a procedure for generating the w vector for each pattern from the appropriate multivariate predictive distribution.

Schaffer's algorithm is written for the special case of estimation and imputation for item missing data. Section II.B showed that the observed responses to the bracketing-type questions produce two classes of data. For reports of actual amounts, the value of the observed coarsened variable, y, is the value of the underlying variable, x. The complementary set of responses are bracketed responses where y is just an indicator of the censoring interval for x, say I_x. One special case of the latter is complete item missing data. Here we will simply interpret complete item missing data as interval

censoring of ln(x) on the range $(0, +\infty)$, and the modified DA algorithm described below will treat complete missing data accordingly. As noted in Section 3.B, zero values in the multivariate response vector will be treated as censoring on the interval $(-\infty, 0]$ for the multivariate lognormal data.

Clearly, interval censored observations included in the multivariate data vector contain information that should be used in the generation of the imputations, w.

Two approaches can be considered for building the interval censoring information into the draws of w. One approach is to use rejection sampling methods to discard any random draws which do not meet the range constraints imposed by the interval censoring indicators contained in the observed y. This would be very inefficient for the multivariate problem considered here. The second approach is to sample from a restricted (truncated) form of the posterior distribution, $p^{*(i)}(w|\theta^{(i)},y)$ for which the draws are forced to meet the range constraints. Since $p^{*(i)}(w|\theta^{(i)},y)$ is a multivariate normal distribution, Devroye (1986) presents a simple algorithm for such a restricted sampling based on draws of random uniform deviates. [See also Gelfand et al. (1990).]

DA Step 3: Update $p^{i}(\theta|y)$ by taking the simple average of $p(\theta|w^{(k)},y)$ over the m sample values of $w^{(k)}$.

$$p^{(i+1)}(\theta \mid y) = m^{-1}\sum_{k=1}^{m} p(\theta \mid w^{(k)},y).$$

The DA algorithm iterates through this 3-step sequence until the desired level of convergence is attained.

4.B The Monte Carlo EM Algorithm

A second imputation-based approach to estimation and inference for the coarsened, bracketed response data is the Monte Carlo EM (MCEM) procedure proposed by Wei and Tanner (1991). As with the standard E-M algorithm, Wei and Tanner's algorithm uses an iterative sequence of expectation (E) and maximization (M) steps to compute the maximizer, θ_{max}, of the posterior likelihood, $p(\theta|y)$.

The E step of the ith iteration of the conventional EM algorithm involves solving the integral equation for the expectation of the log posterior, $\ln(p(\theta|y,w))$:

$$Q_i(\theta,\theta^{(i)}) = \int_W \ln(p(\theta \mid y,w))p(w \mid \theta^{(i)},y)dw$$

The MCEM algorithm uses the Monte Carlo method to approximate this expectation. Using the current value of θ_{max} as the working value, $\theta^{(i)}$, a sample of m observations is generated from $p(w|\theta^{(i)},y)$. The current approximation to the expectation of the log posterior is then computed by the simple averaging (mixing):

$$Q_{i+1}(\theta,\theta^{(i)}) = \frac{1}{m}\sum_{j=1}^{m} \log(p(\theta \mid w^{(K)},y)).$$

The M-step of each iteration involves solving for $\theta_{max}^{(i)}$ by maximizing this mixture of log posteriors with respect to θ. For multi-dimensional maximization problems, Wei and Tanner suggest the use of conjugate gradient or quasi-Newton methods.

Like the DA algorithm, at each iteration the MCEM procedure augments the observed data with multiple imputations for each coarsened observation $y=I_j^x$. Unlike the DA algorithm where the imputed w's are drawn from $p^{(i)}(w|y,\theta^{(i)})$ with $\theta^{(i)}$ a random draw from $p^{(i)}(\theta|y)$, MCEM's draws are made from the predictive distribution, $p^{(i)}(w|y,\theta_{max})$. MCEM does not calculate the full posterior $p(\theta|y)$, only its maximizer. The proposed procedures for handling the interval censored measurement in a modified DA algorithm and non-ignorable coarsening (Section 4.C) could also be used in conjunction with the MCEM algorithm.

4.C Modifying the DA Algorithm to Model/Simulate Nonignorable Coarsening

The posterior likelihood, $p(\theta|y)$, defined in Section 4.A assumes that the coarsening mechanism is ignorable. Data presented in Section 2.D suggest that the bracketed response data may be subject to nonignorable coarsening. Therefore, for purposes of simulation and sensitivity analyses it is useful to extend the DA procedure to explicitly incorporate a model of nonignorable coarsening. For such simulations and sensitivity analyses, it is assumed that the parameters of the coarsening model are specified and not of analytic interest. (In actuality, the nature of the data from the full bracketed response question sequence would permit simultaneous estimation of the coarsening model parameters.)

Returning to the theoretical development of the DA algorithm, the complete integral equation for the posterior when the coarsening mechanism is not ignorable could be written as:

$$p_c(\theta \mid y) = \int K_c(\theta,\phi)g(\phi)\,d\phi,$$

where:

$$K_c(\theta,\phi) = \int_G\int_W p(\theta \mid w,y,g)p(w \mid \phi,y,g)h(g)\,dw\,dg$$

Based on this formulation, $K_c(\theta,\phi)$ is the expectation of the $K(\theta,\phi|g)$ over the stochastic coarsening variate, g. Extending without detailed proof the theoretical development of DA presented by Tanner and Wong, coarsened data requires a DA imputation step which samples not from the predictive distribution $p(w|y)$ but

from the joint predictive distribution $p(w,g|y)$. The "imputation" step (Step 2 in Section 4.A) consists of drawing a sample of m vector pairs (W,G) .

The sampling importance resampling (SIR) procedure described by Rubin(1988) provides a practical approach for generating samples from this complex posterior. Consider decomposing the complex joint distribution:

$$p(w,g|y) = p(w|y)p(g|w,y)$$

As in Step 2 of Section 4.A, generate a sample of $M >> m$ values of w from $p(w|y)$. For each draw compute the ratio:

$$i(w^j) = \frac{p(w,g|y)}{p(w|y)} = p(g|w,y).$$

Next, select a sample of m from the M with probability proportionate to size (PPS) where the measure of size for each of the $j=1,\ldots,M$ original draws is the importance ratio, $i(w^j)$.

For a simple model of univariate coarsening with dichotomous outcomes, $G=0$ (not coarsened) and $G=1$ (coarsened), the m draws are needed only for observations where the observed variate is the coarsened interval response, $Y=I_x$. The only acceptable pairs are of the form $(w=W,\ g=1)$. Therefore the importance ratio that is used to determine the final selection probability for the m draws is:

$$i(w^j) = p(g=1|w=W,y=Y).$$

$$e.g.,\ i(w^j) = \Phi(\gamma_0+\gamma_1 W) \textit{ for the univariate case.}$$

Step 3 of the modified DA algorithm is identical to that described in Section 4.A.

5. Summary

Bracketed response question sequences have proved to be highly effective in reducing the amount of complete item missing data for survey measures of financial variables; however, analysis of these data such as multivariate modeling remains complicated due to the coarsened nature of the observations. Heeringa(1993) suggested the use of newly developed GIBS methods to develop a multiply imputed data set that would enable analysts to conduct multivariate analysis of these data using conventional software systems. The present paper has shifted the emphasis from the multiple imputation of item missing data and interval censored responses, to estimation of parameters under a complete model for the coarsened, bracketed response data. The GIBS data augmentation approach proposed in Section 5 addresses the major shortcomings of the general-location model method presented in the 1993 paper.

It should be noted that the data augmentation algorithm described here lends itself readily to performing multiple imputations for coarsened and completely missing responses. Upon convergence of the algorithm, multiple imputations are obtained by making m draws from $p(w|y,\theta^{(final)})$.

Presently, programming of the modified data augmentation algorithm is underway using the S-Plus system supplemented with subroutines written in lower level languages (Fortran and C+). Upon completion of the programming steps, a large scale simulation study is planned. The simulation study will serve not only to test the accuracy and speed of the program but will examine the performance of the method over a range of data set properties including: 1) degree of coarsening (including complete missing data); 2) variance-covariance of the underlying data; 3) sample size; and 4) various stochastic models for the coarsening mechanism.

Bibliography

Aigner, D.J., & Goldberger, A..S (1970). "Estimation of Pareto's law from grouped observations," *Journal of the American Statistical Association*, Vol. 65, pp. 712-723.

Curtin, R.T., Juster, F.T., & Morgan, J.N. (1989). "Survey estimates of wealth: An assessment of quality," *in* R.E. Lipsig & H.S. Tiu (eds.), *The Measurement of Saving, Investment and Wealth*. Chicago: The University of Chicago Press.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer-Verlag.

Gelfand, A.E., Hills, S., Racine-Poon, A.., & Smith, A.F.M. (1990), "Illustration of Bayesian inference in normal data models using Gibbs sampling," *Journal of the American Statistical Association*, Vol. 85, No. 412, pp. 972-985.

Heeringa, S.G. (1993). "Imputation of item missing data in the Health and Retirement Survey," *The Proceedings of the Section on Survey Methods*, American Statistical Association, pp. 107-116.

Heitjan, D.F. (1989). "Inference from grouped continuous data: A review (with discussion)," *Statistical Science*, Vol. 4, pp. 164-183.

Heitjan, D.F. (1993). "Ignorability and coarse data: Some biomedical examples," *Biometrics*, Vol. 49, pp. 1099-1109.

Heitjan, D.F., & Rubin, D.B. (1990). "Inferences from coarse data via multiple imputation: Age heaping in a Third World nutrition study," *Journal of the American Statistical Association*, Vol. 85, pp. 304-314.

Heitjan, D.F., & Rubin, D.B. (1991). "Ignorability and coarse data," *Annals of Statistics*, Vol. 19, pp. 2244-2253.

Juster, F.T., & Smith, J.P. (1994). "Improving the quality of economic data: Lessons from the HRS." HRS Working Paper Series #94-027, presented at NBER Summer Institute on Health and Aging. Cambridge (MA), July, 1994.

Little, R.J.A. (1993). "Discussion of Heeringa (1993),"

Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 117-119.

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

Little, R.J.A., & Su, H.L. (1987), "Missing data adjustments for partially scaled variables," *Proceedings of the Section on Survey Research Methods*, pp. 644-649.

Maddala, G.S. (1983). *Limited Dependent and Qualitative Variables in Econometrics.* Economic Society Monographs.

Rubin, D. B. (1976). "Inference and missing data," *Biometrika* Vol. 63, pp. 581-592.

Rubin, D.B. (1988). "Using the SIR algorithm to simulate posterior distributions." *In* J.M. Bernardo, M.H. DeGroot, D.V. Lindley, & A.F.M. Smith (eds.), *Bayesian Statistics*, Vol. 3. Oxford University Press, pp. 395-402.

Schaffer, J.L. (1991). "Algorithms for multiple imputation and posterior simulation from incomplete multivariate data with ignorable nonresponse." Unpublished Ph.D. thesis, Department of Statistics, Harvard University.

Tanner, M.A., & Wong, W.H. (1987b). "The calculation of posterior distributions by data augmentation (with discussion)," *Journal of the American Statistical Association*, Vol. 82, pp. 528-540.

Tobin, J. (1958). "Estimation of relationships for limited dependent variables," *Econometrica* Vol. 26, pp. 24-36.

Wachter, K.W., & Trussell, J. (1982). "Estimating historical heights (with discussion)," *Journal of the American Statistical Association*, Vol. 77, pp. 279-303.

Wei, C.G., & Tanner, M.A. (1990), "A Monte Carlo implementation of the EM algorithm and the Poor Man's Data Augmentation algorithm," *Journal of the American Statistical Association*, Vol. 85, no. 411, pp. 699-704.

Notes

[1]The bracketed value category includes cases in which, due to nonresponse or uncertainty, the boundary values for the amount may span two or three of the actual bracket ranges for the item question.

[2]The number of brackets and the associated dollar amounts vary to reflect differences in the properties of the underlying asset distribution.

[3]We could also add "...how to impute..." to this sentence.