Modeling and imputation of semicontinuous survey variables

Joseph L. Schafer Maren K. Olsen *

September 1999

Semicontinuous variables have a proportion of responses at some fixed value and a continuous distribution among the remaining responses. Variables of this type occur in economic surveys of individuals or establishments (e.g. specific types of income or expenditures) where distributions are frequently characterized by a mixture of zeros and continuously distributed positive numbers. In this paper, we review strategies for joint statistical modeling and imputation of semicontinuous survey variables. Algorithms are presented for parameter estimation in the presence of unit and item nonresponse and for the imputation of missing values. Methods and software are demonstrated on data from a variety of sources including the Consumer Expenditures Survey.

Key Words: two-part regression, selection model, Tobit model, survey nonresponse

*Joseph L. Schafer is Associate Professor and Maren K. Olsen is Research Assistant, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802. This research was supported in part by grant 1-P50-DA10075 from the National Institute on Drug Abuse, and by the 1997–98 ASA/NSF/BLS Senior Research Fellow program.

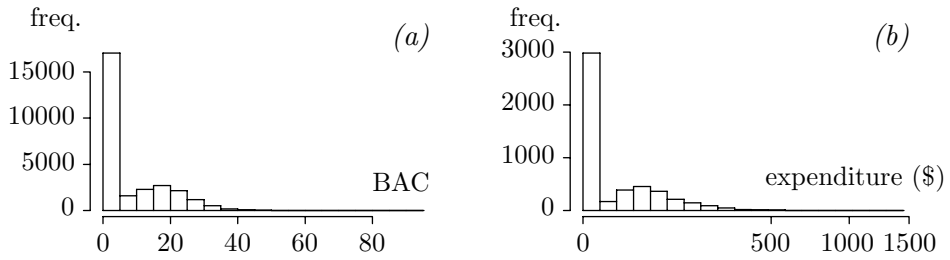

Figure 1: Frequency histograms for (a) values of BAC from the 1993 Fatality Analysis Reporting System, and (b) dollar expenditures on footwear reported in the Consumer Expenditures Survey during the first quarter of 1996.

1 Introduction

In this article we discuss variables with a peculiar type of distribution which arises frequently in a variety of contexts. Consider the two histograms shown in Figure 1 (a) and (b). Histogram (a) displays test results for blood alcohol content (BAC) from 27,633 motor vehicle drivers and pedestrians actively involved in fatal accidents on U.S. highways during 1993. These values were drawn from the Fatality Analysis Reporting System, a data registry maintained by the National Highway Traffic Safety Administration. For 57.3% of the cases shown in Figure 1 (a), the BAC value is zero, indicating no measurable alcohol was found; the remaining values range from 1 to 94, representing blood alcohol levels from 0.01 to 0.94 grams per deciliter (g/dl). The histogram in Figure 1 (b) displays expenditures on footwear reported in the first quarter of 1996 by 4,876 consumer units in the Consumer Expenditure Survey conducted by the U.S. Bureau of Labor Statistics. A consumer unit is a family or group of individuals who pool income and make joint expenditures. To reduce apparent skewness, the expenditures have been plotted on a square-root scale. Nearly 61% of the observed values are zero, and the rest are distributed between $1 and $1,315.

In each of these examples, the quantity of interest is semicontinuous, a mixture of zeros and continuously distributed positive values. Semicontinuous variables differ from variables that are left-censored or truncated in that the zeros are bonafide valid data values rather than proxies for negative or missing responses. Semicontinuous variables pose interest-

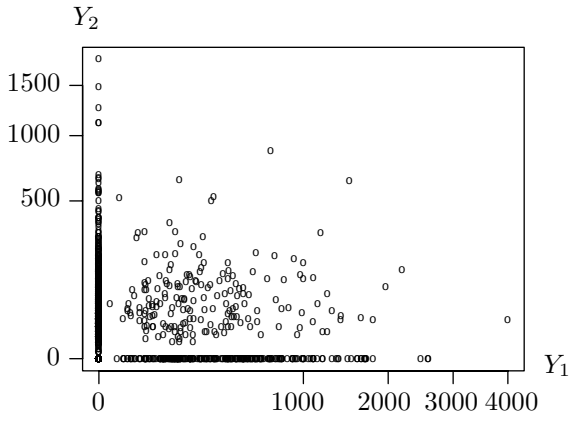

Figure 2: Plot of Y_1 = dollar expenditure on major household appliances versus Y_2 = expenditure on small appliances reported in the Consumer Expenditures Survey during the first quarter of 1996.

ing challenges for data analysts who must assess their relationships with other variables, because these relationships may be quite complicated.

To illustrate the types of relationships that may be found, Figure 2 plots two other variables from the same quarter of the Consumer Expenditure Survey: Y_1 = expenditures on major household appliances (e.g. refrigerators) and Y_2 = expenditures on small household appliances, both plotted on the square-root scale. Among the 4,876 consumer units surveyed, 70.3% had no expenditure in either category ($Y_1 = Y_2 = 0$), 5.3% purchased major appliances but no small appliances ($Y_1 > 0$, $Y_2 = 0$), 20.5% purchased small appliances but no major appliances ($Y_1 = 0$, $Y_2 > 0$), and 3.9% purchased both ($Y_1 > 0$, $Y_2 > 0$). The binary indicators for $Y_1 > 0$ and $Y_2 > 0$ are strongly related; those who purchased major appliances were nearly twice as likely to purchase small appliances as those who did not. But further examination reveals additional, more subtle relationships as well. Among those who purchased small appliances ($Y_2 > 0$), the correlation between $Y_2^{1/2}$ and the binary indicator for $Y_1 > 0$ is $r = +0.11$; the amount spent on small appliances bears a significant positive relationship to the purchase of major appliances. But among those who purchased major appliances ($Y_1 > 0$), the the amount spent on major appliances $Y_1^{1/2}$ is significantly

negatively correlated with the binary indicator for small appliance purchase ($r = -0.12$). Finally, among those who purchased in both categories ($Y_1 > 0$, $Y_2 > 0$), the correlation between $Y_1^{1/2}$ and $Y_2^{1/2}$ is essentially zero ($r = +0.03$).

Semicontinuous variables also challenge primary data collectors who may need to impute values of these variables for nonrespondents. An imputation procedure not specifically tailored to semicontinuous data may seriously distort the variable's marginal distribution or its relationships with other quantities. A sensible imputation procedure should be capable of preserving distributional shapes as shown in Figure 1 and relationships as shown in Figure 2. A variety of new model-based imputation procedures, including multiple imputation (Rubin, 1987), have been developed for multivariate continuous and categorical data (Schafer, 1997). To date, however, little has been written on the unique problem of imputing semicontinuous data.

In the remainder of this article, we review and critique a variety of methods for statistical modeling and model-based imputation of semicontinuous variables. When discussing the merits of a model, it is important to keep in mind the purposes for which the model will be used. If the goal is to analyze data for scientific understanding of populations and data-generating mechanisms, then interpretation of model parameters becomes vitally important, and one should specify a model whose parameters are meaningful and relevant to the scientific issues at hand. If the model is merely a device for producing imputations, however, then interpretability of parameters is less crucial; the model should then be judged according to its ability to fit the data well and to generate plausible simulated values for the missing observations. We shall see that in some cases an imputation model may perform well even if it contains parameters that cannot be estimated at all.

In Section 2 we present models for a single semicontinuous variable without covariates. This univariate problem, albeit simplistic, lays the foundation for models and imputation procedures in more complicated situations, including: one semicontinuous variable with

additional categorical and continuous covariates (Section 3); one semicontinuous variable measured repeatedly over time (Section 4); and the general multivariate problem where an arbitrary number of continuous, categorical, and semicontinuous variables may need to be modeled and imputed jointly (Section 5). In some respects, the general multivariate problem is still unsolved. We do not yet have a workable, self-consistent procedure for modeling and imputing large numbers of semicontinuous variables at once, but several approximate methods are available; in particular, a multivariate procedure based on a normal model appears to work quite well. We conclude in Section 6 with a discussion of ongoing efforts and future directions in this important and challenging area.

2 Models for one semicontinuous variable

2.1 The Tobit model

The Tobit model (Tobin, 1958; Amemiya, 1984) was devised for a truncated or limited dependent variable taking values on the positive real line $(0, \infty)$, where negative values have been censored and coded as zero. Suppose a sample $z_1, \ldots, z_n$ is drawn from a normal population with mean μ and variance σ^2, but we only observe $y_1, \ldots, y_n$ where $y_i = \max(z_i, 0)$. More generally, one could include covariates by setting $E(z_i) = x_i^T \beta$, but this would add little to the present discussion so we will simply keep $E(z_i) = \mu$ for $i = 1, \ldots, n$. Maximum-likelihood (ML) estimates for $\theta = (\mu, \sigma^2)$ may be found by an EM algorithm as shown by Little and Rubin (1987, ch. 11). The E-step requires calculation of $S_z = \sum_{i=1}^n E(z_i \mid y_i)$ and $S_{zz} = \sum_{i=1}^n E(z_i^2 \mid y_i)$ under the current estimate of θ. For the non-censored cases ($y_i > 0$) the expectations are simply $E(z_i \mid y_i) = y_i$ and $E(z_i^2 \mid y_i) = y_i^2$; for the censored cases ($y_i = 0$) they are

$$\begin{aligned} E(z_i \mid y_i) &= \mu - \sigma\lambda(-\mu/\sigma), \\ E(z_i^2 \mid y_i) &= \sigma^2 + \mu^2 - \mu\sigma(-\mu/\sigma), \end{aligned}$$

where $\lambda(z) = \phi(z)/\Phi(z)$ is the inverse of the Mills ratio, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and cumulative distribution functions, respectively. The M-step updates the parameter estimates for μ and σ^2, replacing them by $\hat{\mu} = n^{-1}S_z$ and $\hat{\sigma}^2 n^{-1}S_{zz} - (n^{-1}S_z)^2$. E- and M-steps are performed alternately until the estimates of μ and σ^2 stabilize.

From a theoretical standpoint, application of the Tobit model to semicontinuous variables can be problematic. When $y_1, \ldots, y_n$ are truly semicontinuous, $y_i = 0$ is a valid, self-representing data point rather than a mask for an unseen negative value. The underlying sample $z_1, \ldots, z_n$ does not really exist, and interpretation of μ and σ^2 becomes dubious. It is not clear, for example, what these parameters would mean if the Tobit model were applied to the data of Figure 1 (a) or (b), because negative values of blood alcohol content or negative dollar expenditures are not conceivable.

Aside from these difficulties of interpretation, Tobit models are often unattractive simply because they do not fit. Under the Tobit model, the probability mass at zero is tied to the location and scale of the continuously distributed values as $P(y_i = 0) = \Phi(-\mu/\sigma)$. With real data, this condition tends to be satisfied only by accident. To illustrate, we fit the Tobit model to the square roots of the $n = 4,876$ footwear expenditure values shown in Figure 1 (b). The EM algorithm quickly converged to $\hat{\mu} = -2.21$ and $\hat{\sigma} = 10.3$, yielding an estimated rate of zero expenditures equal to $\Phi(-0.214) = .585$. In the sample, however, the observed rate of zero expenditures is .608; the 95% confidence interval calculated in the standard manner is $(.595, .622)$, which does not even cover the Tobit estimate. With only two parameters, the Tobit model is simply not rich enough to generate the distributional shapes found in many semicontinuous variables.

2.2 Selection models

A more general type of model popular among econometricians is the class of stochastic censoring or self-selection models (Heckman, 1974, 1976). In Heckman's selection model,

the continuous variable of interest is assumed to be censored or missing if an unobserved, normally distributed variable falls below a threshold. A selection model for semicontinuous data can be formulated as follows. Suppose that (z_i, r_i), $i = 1, \ldots, n$ are a sample from a bivariate normal distribution,

$$\begin{pmatrix} z_i \\ r_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu \\ \gamma \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right],$$

but we observe only $y_1, \ldots, y_n$, where

$$y_i = \begin{cases} z_i & \text{if } r_i > 0, \\ 0 & \text{if } r_i \leq 0. \end{cases}$$

The variance of r_i is set equal to 1 for identifiability. The parameter γ, which determines the point mass $P(y_i = 0) = \Phi(-\gamma)$, is not tied to the location or scale of z_i, so this model tends to fit the proportion of zeros better than the Tobit model.

Little and Rubin (1987, ch. 11) describe an EM algorithm for estimating $\theta = (\mu, \gamma, \sigma^2, \rho)$. The E-step calculates expected sufficient statistics $S_z = \sum_i E(z_i \mid y_i)$, $S_r = \sum_i E(r_i \mid y_i)$, $S_{zz} = \sum_i E(z_i^2 \mid y_i)$, $S_{zr} = \sum_i E(z_i r_i \mid y_i)$, and $S_{rr} = \sum_i E(r_i^2 \mid y_i)$ under the current parameter estimates, and the M-step updates the estimates as $\hat{\gamma} = n^{-1} S_r$, $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 \hat{\gamma}$, $\hat{\sigma}^2 = \hat{\psi} + \hat{\beta}_1^2$, and $\hat{\rho} = \hat{\beta}_1 / \hat{\sigma}$, where $\hat{\beta}_0 = n^{-1} S_z - \hat{\gamma}\hat{\beta}_1$,

$$\begin{aligned} \hat{\beta}_1 &= (S_{zr} - n^{-1} S_z S_r)/(S_{rr} - n^{-1} S_r^2), \\ \hat{\psi} &= (n^{-1} S_{zz} - n^{-2} S_z^2) - \hat{\beta}_1 (n^{-1} S_{zr} - n^{-2} S_z S_r). \end{aligned}$$

For the observations of $y_i = 0$, the E-step expectations are

$$\begin{aligned} E(z_i \mid y_i) &= \mu - \rho\sigma\lambda(-\gamma), \\ E(r_i \mid y_i) &= \gamma - \lambda(-\gamma), \\ E(z_i^2 \mid y_i) &= \sigma^2 + \mu^2 - \rho\sigma\lambda(-\gamma)(2\mu - \rho\sigma\gamma), \\ E(z_i r_i \mid y_i) &= \mu(\gamma - \lambda(-\gamma)) + \rho\sigma, \\ E(r_i^2 \mid y_i) &= 1 + \gamma^2 - \gamma\lambda(-\gamma), \end{aligned}$$

where $\lambda(\cdot)$ is the inverse of the Mills ratio. For the observations of $y_i \neq 0$, the expectations become $E(z_i \mid y_i) = y_i$, $E(z_i^2 \mid y_i) = y_i^2$, $E(z_i r_i \mid y_i) = y_i E(r_i \mid y_i)$,

$$\begin{aligned} E(r_i \mid y_i) &= \hat{r}_i + \sqrt{1-\rho^2}\,\lambda\left(\hat{r}_i^2/\sqrt{1-\rho^2}\right), \\ E(r_i^2 \mid y_i) &= (1-\rho^2) + \hat{r}_i^2 + \hat{r}_i\sqrt{1-\rho^2}\,\lambda\left(\hat{r}_i^2/\sqrt{1-\rho^2}\right), \end{aligned}$$

where $\hat{r}_i = \gamma + \rho(y_i - \mu)/\sigma$. These last two expectations correct a small error in the formulas of Little and Rubin (1987), who neglected to condition on the value of $z_i = y_i$ which is observed when $y_i \neq 0$.

We applied this procedure to the square-roots of the footwear expenditure values from Figure 1 (b) and obtained $\hat{\mu} = 1.95$, $\hat{\sigma} = 6.82$, $\hat{\gamma} = -0.278$, and $\hat{\rho} = 0.983$. It is somewhat surprising that the correlation ρ can be estimated at all, because r_i is never directly observed and z_i is only seen when $y_i \neq 0$. In a sense the selection model is over-parameterized because the data contain very little information about ρ. This can be illustrated by examining the loglikelihood function $\sum_i \log L(\theta \mid y_i)$, where

$$L(\theta \mid y_i) = \begin{cases} \Phi(-\gamma) & \text{if } y_i = 0, \\ \frac{1}{\sigma}\,\phi\left(\frac{y_i-\mu}{\sigma}\right)\,\Phi\left(\frac{\gamma+\rho(y_i-\mu)/\sigma}{\sqrt{1-\rho^2}}\right) & \text{if } y_i \neq 0. \end{cases}$$

Figure 3 shows a plot of the profile loglikelihood function for ρ, the loglikelihood obtained by fixing ρ at any specific value and maximizing with respect to the other three parameters. An approximate 95% confidence interval for ρ, consisting of all values for which the profile loglikelihood multiplied by two lies within 4 units of its maximum, ranges from 0.977 to 0.988. Appealing to the well-known result that $z = .5\log((1+r)/(1-r))$ is approximately normally distributed with variance $1/(n_0 - 3)$ if r is a sample correlation coefficient from a sample of size n_0, it appears that these $n = 4,876$ values of y_i carry information about ρ equivalent to only about $n_0 \approx 160$ observations of (z_i, r_i). This high rate of missing information about ρ causes the EM algorithm to converge very slowly. Another feature of the selection model, as pointed out by Little and Rubin (1987), is that the estimate of ρ can be extremely sensitive to distributional shape; virtually all the information about this

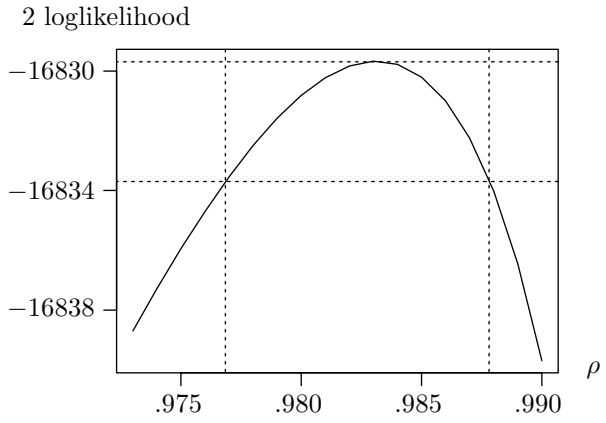

Figure 3: Profile loglikelihood function for ρ, with limits of approximate 95% confidence interval.

parameter comes from sample skewness. In this example, taking logs rather than square roots of the non-zero values of y_i causes the estimated value of ρ to drop almost to zero.

Finally, when applied to semicontinuous data, selection models share with the Tobit model the same difficulties of parameter interpretation. If $y_i = 0$ represents a valid data point rather than a proxy for some underlying nonzero value, then μ and σ^2 do not describe the mean and variance of any real population of interest.

2.3 A two-part model

A simpler way to model semicontinuous data is to present them as a two-part mixture of a normal distribution and a degenerate point mass. Suppose that the semicontinuous observations $y_1, \ldots, y_n$ are recoded as two variables (w_i, z_i), $i = 1, \ldots, n$, where

$$w_i = \begin{cases} 1 & \text{if } y_i \neq 0, \\ 0 & \text{if } y_i = 0, \end{cases} \tag{1}$$

$$z_i = \begin{cases} g(y_i) & \text{if } y_i \neq 0, \\ \text{irrelevant} & \text{if } y_i = 0, \end{cases} \tag{2}$$

and g is some monotonically increasing function (e.g. log) chosen to make the nonzero values of y_i approximately normally distributed. The binary indicators w_i are assumed to be Bernoulli with $P(w_i = 1) = \pi$, and the conditional distribution of z_i when $w_i = 1$ is

assumed to be $N(\mu, \sigma^2)$; when $w_i = 0$ the variable z_i becomes irrelevant and is not modeled. Under these assumptions, the likelihood for $\theta = (\pi, \mu, \sigma^2)$ factors into distinct functions for π and (μ, σ^2),

$$L(\theta) \propto \prod_{i=1}^{n} \pi_i^{w_i}(1-\pi_i)^{1-w_i} \prod_{i:\, y_i \neq 0} \sigma^{-1} \exp\left\{ \frac{(z_i-\mu)^2}{2\sigma^2} \right\}, \tag{3}$$

and ML estimates are available in closed form as $\hat{\pi} = n^{-1} \sum_{i=1}^{n} w_i$,

$$\begin{aligned} \hat{\mu} &= n_1^{-1} \sum_{i:\, y_i \neq 0} z_i, \\ \hat{\sigma}^2 &= n_1^{-1} \sum_{i:\, y_i \neq 0} (z_i - \hat{\mu})^2, \end{aligned}$$

where $n_1 = \sum_{i=1}^{n} w_i$.

Note that the parameters μ and σ^2 of this two-part model have natural meaning as the conditional mean and variance of $z_i = g(y_i)$ given that $y_i \neq 0$. The simplicity of estimation and natural interpretation make two-part models attractive for data analysis. In one of the earliest applications of a two-part model, Manning, Morris, and Newhouse (1981) and Duan et al. (1983) described medical expenditures by a pair of regression functions, one for the probability of expenditure, the other for the mean of the log-expenditure if there was any.

In two-part modeling, it is sometimes convenient to model z_i for all cases regardless of whether $y_i = 0$ or $y_i \neq 0$, and then apply missing-data procedures or algorithms such as EM to handle the unobserved values of z_i for the $y_i = 0$ cases. Suppose that we model the (w_i, z_i) pair as

$$w_i \sim \text{Bernoulli}(\pi), \tag{4}$$

$$z_i \mid w_i = 1 \sim N(\mu, \sigma^2), \tag{5}$$

$$z_i \mid w_i = 0 \sim N(\mu^*, \sigma^2), \tag{6}$$

and regard z_i as missing whenever $w_i = 0$. This nonresponse mechanism is ignorable in the sense defined by Little and Rubin (1987) because missingness on z_i is determined by

w_i which is always observed. The observed-data likelihood function for the augmented set of parameters $\theta^* = (\pi, \mu, \sigma^2, \mu^*)$, which is obtained by integrating the missing observations out of the complete-data likelihood, is

$$L(\theta^*) \propto \prod_{i=1}^{n} \pi_i^{w_i}(1-\pi_i)^{1-w_i} \prod_{i:\, y_i \neq 0} \sigma^{-1} \exp\left\{\frac{(z_i-\mu)^2}{2\sigma^2}\right\} \times \int \prod_{i:\, y_i = 0} \sigma^{-1} \exp\left\{\frac{(z_i-\mu^*)^2}{2\sigma^2}\right\} dZ_{mis}, \quad (7)$$

where $Z_{mis} = \{z_i : y_i = 0\}$. This function reduces to (3) because the integral is a constant. Thus (4)–(6) is equivalent to the two-part model except for the new parameter μ^* which does not appear in the likelihood and is therefore inestimable.

The formulation (4)–(6) is useful because it is a special case of the *general location model*, a model for multivariate data containing both categorical and continuous variables (Little and Schluchter, 1985). Incomplete-data procedures for this model, including EM-type algorithms for model fitting and Markov chain Monte Carlo algorithms for multiple imputation, are already available (Schafer, 1997, ch. 9). These procedures can be adapted to semicontinuous observations $y_1, \ldots, y_n$ by recoding them as (w_i, z_i), $i = 1, \ldots, n$ where z_i is missing whenever $w_i = 0$. The pair (w_i, z_i) can be imputed jointly and values for y_i derived by the following rule: Set $y_i = g^{-1}(z_i)$ if $w_i = 1$ and $y_i = 0$ if $w_i = 0$. Depending on the procedure, the presence of an extra inestimable parameter μ^* may not be a problem at all. For example, an EM algorithm applied to these data will converge to a unique estimate for $\theta = (\pi, \mu, \sigma^2)$ and a non-unique but irrelevant value for μ^*. If this lack of identification were troublesome, we could impose the constraint $\mu^* = \mu$ which would have no effect on the relevant parts of the model, but such constraints may in fact be unnecessary. The behavior of algorithms for the general location model, with constraints and with inestimable parameters, will be explored in Sections 3 and 5.

Finally, suppose that we regard the binary indicator w_i as a recoded version of an unseen variable r_i assumed to be normally distributed with mean γ and variance 1, where $w_i = 0$

if $r_i \leq 0$ and $w_i = 1$ if $r_i > 0$. If r_i is assumed to be independent of z_i, then the two-part model can be expressed as

$$\begin{pmatrix} z_i \\ r_i \end{pmatrix} \sim N\left[\begin{pmatrix} \mu \\ \gamma \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix} \right],$$

where $\pi = \Phi(\gamma)$. Thus the two-part model can be regarded as a special case of the selection model where the troublesome parameter ρ has been set to zero. By independence, μ and σ^2 are the mean and variance of z_i both marginally and conditionally given r_i, so the difficulties of parameter interpretation found with the selection model are no longer present.

3 One semicontinuous variable with covariates

3.1 Covariates completely observed

In this section, we investigate models and imputation procedures for one semicontinuous variable with additional covariates. When the covariates have no missing values, procedures can be developed by straightforward extension of the two-part model of Section 2.3. Suppose that Y is a semicontinuous variable to be modeled and perhaps imputed given covariates $X_1, \ldots, X_r$. We will suppose that Y is observed and zero for sample units $1, \ldots, n_1$, observed and nonzero for units $n_1 + 1, \ldots, n$, and missing for units $n + 1, \ldots, m$. Nonresponse on Y is assumed to be ignorable, which means that the probabilities of missingness may depend on the $X_1, \ldots, X_r$ but are not directly related to Y. A diagram of the sample is shown in Figure 4.

Suppose that we recode Y into a binary indicator W and a continuous variable Z defined as in (1)–(2). These recodes are displayed in Figure 4 with an identity transformation used for g. As before, we model W as Bernoulli and Z as conditionally normal given $W = 1$, but we now introduce dependence on the covariates through standard logistic and linear regression. Let (w_i, z_i) and x_i^T denote the sample values of (W, Z) and $(X_1, \ldots, X_r)^T$, respectively, for unit i. The model is $w_i \sim \text{Bernoulli}(\pi_i)$ and $z_i \mid (w_i = 1) \sim N(\mu_i, \sigma^2)$,

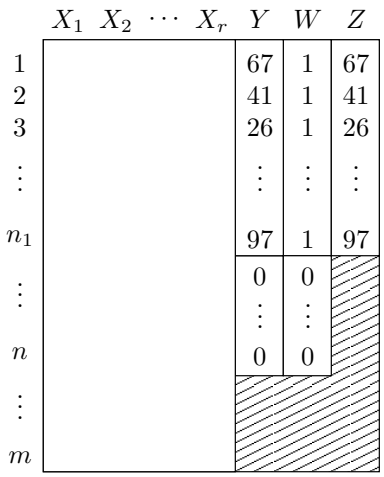

Figure 4: One semicontinuous variable Y recoded as a binary indicator W and a continuous variable Z, with completely observed covariates $X_1, \ldots, X_r$.

where $\pi_i = \exp(x_i^T\beta)/[1+\exp(x_i^T\beta)]$ and $\mu_i = x_i^T\gamma$. The likelihood function for this model is identical to (3) except that π and μ are replaced by π_i and μ_i. Because the likelihood factors into distinct functions pertaining to β and (γ, σ^2), parameters can be estimated in two steps: first, fit the logistic regression of W on $X_1, \ldots, X_r$ using sample units $1, \ldots, n$; second, fit the linear regression of Z on $X_1, \ldots, X_r$ using sample units $1, \ldots, n_1$.

Two-part regression has been applied in econometric analyses for nearly two decades (Manning, Morris, and Newhouse, 1981; Duan et al., 1983). Similar models for excess zeros in count data have been developed by Heilbron (1989) and Lambert (1992). For simplicity, we have supposed that the same covariates $X_1, \ldots, X_r$ appear as predictors in the logistic and linear parts, but this is not necessary; overlapping or distinct sets of predictors may also be used. It is quite conceivable that a predictor whose coefficient is positive in one part may have a zero or negative coefficient in the other part; effects of this type were seen in Figure 2.

If one had to impute the missing values of Y for cases $n+1, \ldots, m$ as shown in Figure 4, multiple imputations could be created as follows.

1. Fit the logistic regression model for W using cases $1, \ldots, n$, saving the ML estimates

$\hat{\beta}$ and their estimated covariance matrix $\hat{V}(\hat{\beta})$.

2. Draw a new random set of coefficients β^* from a multivariate normal distribution centered at $\hat{\beta}$ with covariance matrix $\hat{V}(\hat{\beta})$.

3. Draw w_i from a Bernoulli distribution with probability $\pi_i^* = \exp(x_i^T\beta^*)/[1+\exp(x_i^T\beta^*)]$ independently for $i = n+1, \ldots, m$.

4. Fit the linear regression for Z using cases $1, \ldots, n_1$, saving the least-squares estimates $\hat{\gamma}$, the vector of residuals $\hat{\epsilon}$ with elements $\hat{\epsilon}_i = z_i - x_i^T\hat{\gamma}$ for $1, \ldots, n_1$, and the matrix $(X^TX)^{-1}$ where $X = (x_1, \ldots, x_{n_1})^T$.

5. Draw a random value of σ^2 as $\sigma^{2*} = \hat{\epsilon}^T\hat{\epsilon}/A$, where A is a chisquare variate with $n_1 - r$ degrees of freedom, and then draw γ^* from a multivariate normal distribution centered at $\hat{\gamma}$ with covariance matrix $\sigma^{2*}(X^TX)^{-1}$.

6. Draw z_i from a normal distribution with mean $\mu_i^* = x_i^T\gamma^*$ and variance σ^{2*} independently for $i = n+1, \ldots, m$.

7. Set $y_i = 0$ if $w_i = 0$ and $y_i = g^{-1}(z_i)$ if $w_i = 1$ for $i = n+1, \ldots, m$.

Repeating these steps M times produces a set of M multiple imputations of $Y_{mis} = (y_{n+1}, \ldots, y_m)$ drawn from an approximate Bayesian predictive distribution under the two-part model. Note that the model fits in steps 1 and 4 would not change and need only be performed once.

3.2 Covariates with missing values

If the covariates $X_1, \ldots, X_r$ have occasional missing values, then a pair of regression models may no longer suffice; the covariates may need to be jointly modeled and perhaps imputed along with the semicontinuous variable Y. A framework for joint modeling and imputation of incompletely observed categorical and continuous variables is provided by the general

location model (Little and Schluchter, 1985; Schafer, 1997, ch. 9). When our covariates are categorical—or when they can be treated as categorical for purposes of imputation—then a general location model that jointly describes $X_1, \ldots, X_r$ and Y is easily constructed.

A general location model combines (a) a loglinear model for describing relationships among categorical variables with (b) a multivariate linear regression for describing the correlations among continuous variables and their relationships to the categorical ones. Let $(W_1, \ldots, W_p)$ and $(Z_1, \ldots, Z_q)$ denote a group of categorical variables and a group of continuous variables, respectively, collected for a sample of n subjects. If W_j takes d_j distinct values, then each subject belongs to a unique cell of the $W_1 \times \cdots \times W_p$ contingency table whose total number of cells is $D = \prod_{j=1}^{p} d_j$. Let $\pi = (\pi_1, \ldots, \pi_D)$ denote the probabilities associated with these cells. These probabilities are assumed to follow a standard loglinear model in which the log-probabilities are expressed as the sum of an intercept, main effects for $W_1, \ldots, W_p$, and possibly interactions among them. When formulating a loglinear model, the user designates a set of associations among $W_1, \ldots, W_p$ for the model to preserve—for example, (W_1, W_2) and (W_2, W_3) produces a model in which W_1 is related to W_2, and W_2 is related to W_3, but W_1 and W_3 are conditionally independent given W_2. Methods for fitting and interpreting loglinear models are reviewed by Bishop, Fienberg and Holland (1975) and Agresti (1990).

The second part of a general location model is a multivariate regression for $Z_1, \ldots, Z_q$ given $W_1, \ldots, W_p$. Let z_i be the realized value of $(Z_1, \ldots, Z_q)^T$ for subject i and $Z = (z_1, \ldots, z_n)^T$ the $n \times q$ matrix of continuous data for the entire sample. Let U be a $n \times D$ matrix of dummy indicators which records the cell locations $1, \ldots, D$ in the contingency table for each subject. The (i, j)th element of U is equal to 1 if subject i belongs to cell j and 0 otherwise. The regression model is

$$Z = UA\beta + \epsilon, \tag{8}$$

where A is a $D \times s$ design matrix, β is an $s \times q$ matrix of regression coefficients, and ϵ is an

$n \times q$ matrix of residuals whose rows are independent and normally distributed with mean 0 and covariance matrix Σ. The design matrix A, which must be specified by the user, defines the relationship between each continuous variable Z_j and the cells of the $W_1 \times \cdots W_p$ table. The columns of A typically include a constant term for an intercept and dummy codes or contrasts defining the main effects of $W_1, \ldots, W_p$ and the interactions among them. The regression coefficients for predicting Z_j are found in the jth column of β.

In econometric texts, a regression model of the form (8) is sometimes called a *standard multivariate regression* because the same set of regressors predicts each column of the response matrix Z. If a certain interaction among $W_1, \ldots, W_p$ is allowed to influence one of the response variables Z_1, then it must be allowed to influence $Z_2, \ldots, Z_q$ as well. More generally, one could apply a different set of regressors to each response variable. This more complicated model, which is sometimes called *seemingly unrelated regression*, can be expressed in the form (8) if we take the columns of A to be the union of the regressors for $Z_1, \ldots, Z_q$ and constrain some elements of β to be fixed at zero a priori. The estimation and imputation procedures for the general location model described by Schafer (1997, ch. 9) were designed for regressions of the standard type, not the seemingly unrelated type, because a priori constraints on the elements of β are not allowed. This detail becomes important when the general location model is applied to semicontinuous data, as we shall now see.

Returning to the situation shown in Figure 4, suppose that we have one semicontinuous variable Y and covariates $X_1, \ldots, X_r$ possibly having missing values. If $X_1, \ldots, X_r$ are categorical, we can create a general location model as follows: define $W_1 = 1$ if $Y \neq 0$ and $W_1 = 2$ if $Y = 0$; set $W_{j+1} = X_j$ for $j = 1, \ldots, r$; and set $Z_1 = g(Y)$ if $Y \neq 0$ and missing if $Y = 0$. The loglinear model for $W_1, \ldots, W_{r+1}$ should be designed to capture important relationships among these variables. The regression model for Z_1 should reflect important relationships between Z_1 and $W_2, \ldots, W_{r+1}$. But note that it is not possible to estimate an

effect of W_1 on Z_1 because Z_1 is observed whenever $W_1 = 1$ and missing whenever $W_1 = 2$. If the regression design matrix A contains a dummy indicator or contrast corresponding to W_1, then the model will contain an inestimable parameter. If any effect pertaining to W_1 is omitted from A, then under usual circumstances all parameters should be well estimated. By omitting the W_1-effect from the design matrix, we can fit this model and jointly impute missing values for all variables $X_1, \ldots, X_r, Y$ at once using the algorithms described by Schafer (1997, ch. 9).

Now if some of the covariates $X_1, \ldots, X_r$ are continuous, then these may be included in the general location model as $Z_2, Z_3, \ldots$. In general, these continuous variables may be related to the binary indicator W_1, and it would be highly desirable for the regression part of the general location model to reflect these relationships. If an effect for W_1 is coded into the design matrix A, however, then an inestimable relationship between W_1 and Z_1 is introduced. Strategies for handling inestimable parameters will be presented in Section 5 when we discuss the general multivariate semicontinuous problem.

3.3 Example: imputation of missing BAC

Blood alcohol content (BAC), which was displayed in Figure 1 (a), is a key variables compiled in the Fatality Analysis Reporting System (FARS). The National Highway Traffic Safety Administration (NHTSA) defines an accident to be alcohol-related if a driver, pedestrian, or pedacyclist involved had a BAC of 0.01 g/dl or higher. Analyses of BAC levels from FARS figure prominently in policy debates regarding legal limits for drunk driving. In any given year, however, nearly half of the BAC values in FARS are missing. Missing values often occur when police at the accident scene determine, because of the time of day or nature of the accident, that alcohol was probably not a factor. Thus there is reason to believe that the alcohol levels of those who have missing values for BAC are substantially lower, on average, than the observed values for BAC depicted in Figure 1 (a).

Since 1982, NHTSA has imputed missing values for BAC in three broad categories: BAC=0 (no alcohol), $0 < \text{BAC} \leq 9$, and $\text{BAC} \geq 10$. This three-level classification was sufficient for many analyses of interest because levels of 0.10 g/dl or greater corresponded to the legal definition of intoxication in most states. Probabilities of falling into the three classes were estimated by a three-level discriminant model on the basis of other recorded variables found to be significantly related to BAC, including time of day, type of vehicle, age and sex of driver, and a dichotomous variable indicating whether police at the scene believed alcohol to be involved (Klein, 1986).

In recent years, some states have reduced their legal limits to 0.08 or even 0.02 for drivers under the age of 21, and other states are considering doing so. As a result, estimating proportions for alternative class definitions such as $\text{BAC} \geq 8$ has become desirable. The need for greater detail in imputed BAC values led NHTSA to consider alternative imputation procedures that would reflect this variable's semicontinuous nature. Extending the earlier work by Klein (1986), we have developed an extension of the general location model to multiply impute semicontinuous values for BAC. Model fitting and imputation are carried out in S-PLUS (Mathsoft, Inc., 1997) using routines from the CAT and MIX libraries (Schafer, 1996a, 1996b). Using this new method, NHTSA has re-imputed BAC for each year of FARS since 1982, and will shift entirely to the new procedure for 1999 and beyond. A description of the method and preliminary results have been published by NHTSA in the form of a Research Note (NHTSA, 1998).

4 Modeling a semicontinuous longitudinal response

4.1 A two-part random-effects model

Suppose that a semicontinuous response is recorded for a sample of individuals at multiple points in time. Data of this type arise frequently in panel surveys. For example, the data shown in Figure 5 were drawn from the Adolescent Alcohol Prevention Trial, a longitudinal

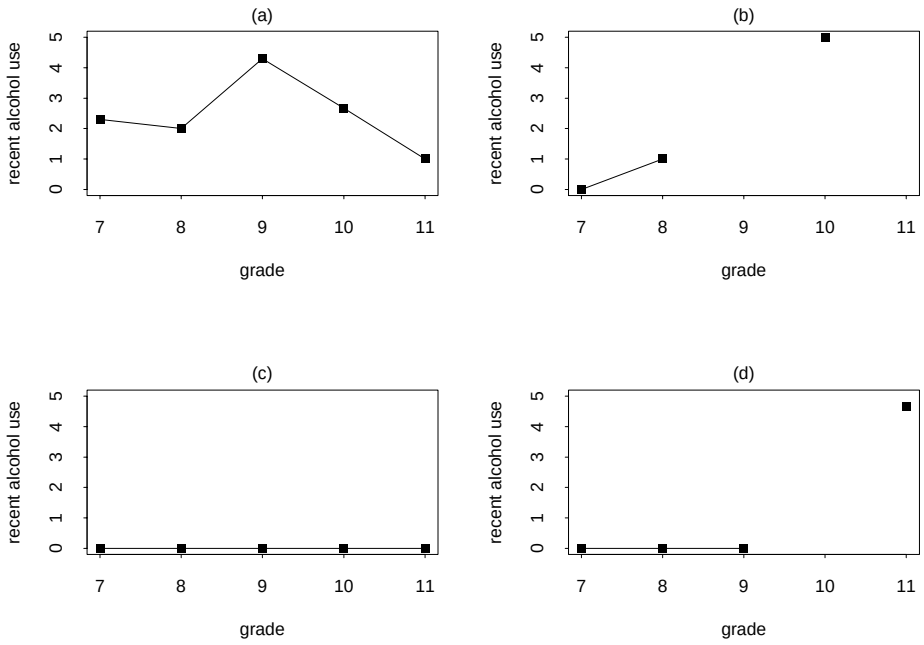

Figure 5: Reported levels of recent alcohol use in grades 7–11 for four subjects in the Adolescent Alcohol Prevention Trial.

study of substance use among students in the Los Angeles metropolitan area (Hansen and Graham, 1991). The four graphs show levels of reported recent alcohol use among four AAPT subjects from grade 7 to grade 11. Subject (a) reports a moderate amount of use in grade 7, a high amount by grade 9, and a low amount by grade 11; subject (c) reports no use at any occasion; and subjects (b) and (d) report a mixture of no use and use with missing values at some time points. Conventional linear models for repeated measures, which can be fit with a variety of software packages including SAS Proc Mixed (Littell *et al.*, 1996), would describe these data poorly because of the preponderance of zeros. Consistent estimation of a population-averaged regression function is possible with generalized estimating equations (GEE) (Zeger, Liang, and Albert, 1988), but a marginal-mean model would also fail to recognize the qualititative distinctions between zero and nonzero responses. Characterizing adolescent substance use with two processes, one binary and one continuous, is theoretically appealing and provides a richer description than a

model with a single mean function.

The two-part regression model described in Section 3.1 can be extended to panel data in the following way. Let Y_{ij} denote the response for subject i at occasion $j = 1, \ldots, n_i$, which we recode as

$$W_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \neq 0, \\ 0 & \text{if } Y_{ij} = 0, \end{cases}$$

and

$$Z_{ij} = \begin{cases} g(Y_{ij}) & \text{if } Y_{ij} \neq 0, \\ \text{irrelevant} & \text{if } Y_{ij} = 0. \end{cases}$$

We propose a pair of correlated random-effects models, one for the logit probability of $W_{ij} = 1$ and one for the mean conditional response $E(Z_{ij} \mid W_{ij} = 1)$. The logit model is

$$\eta_i = X_i\beta + V_i c_i, \tag{9}$$

where $\pi_{ij} = P(W_{ij} = 1)$, η_i is the vector with elements $\eta_{ij} = \log[\pi_{ij}/(1-\pi_{ij})]$, $j = 1, \ldots, n_i$, and X_i $(n_i \times q_c)$ and V_i $(n_i \times p_c)$ are matrices of covariates pertaining to the fixed and random effects, respectively. Times of measurement may be included in X_i and possibly V_i, allowing intercepts, slopes, etc. to vary by subject. The continuous-response model is

$$Z_i = X_i^*\gamma + V_i^* d_i + \epsilon_i, \tag{10}$$

where Z_i is the vector of length n_i^* containing all relevant values of Z_{ij} for subject i, the values corresponding to $W_{ij} = 1$. The residuals ϵ_i are assumed to be distributed as $N(0, \sigma^2 I)$, and X_i^* $(n_i^* \times q_d)$ and V_i^* $(n_i^* \times p_d)$ are matrices of covariates. Finally, the random coefficients from the two parts are assumed to be jointly normal and possibly correlated,

$$b_i = \begin{pmatrix} c_i \\ d_i \end{pmatrix} \sim N\left(0, \psi = \begin{pmatrix} \psi_{cc} & \psi_{cd} \\ \psi_{dc} & \psi_{dd} \end{pmatrix}\right). \tag{11}$$

If subject i reports $Y_{ij} = 0$ at every time point, then $n_i^* = 0$ and the subject does not contribute to the estimation of γ, σ^2, ψ_{dd}, or ψ_{cd}.

The model defined by (9)–(11) is a natural generalization of linear random-effects models first proposed by Hartley and Rao (1967) and popularized through the computational work

of many authors (Laird and Ware, 1982; Jennrich and Schluchter, 1986); Laird, Lange, and Stram, 1987; Lindstrom and Bates, 1988). The same covariates may appear in the logit and linear parts of the model, but this is not required. Intercepts and slopes for either curve can be either fixed or random, and additional static or time-varying covariates may be included in either curve. Responses need not be recorded at the same set of time points for all individuals; the data may be unbalanced by design or have ignorably missing values. Note that if $\psi_{cd} = 0$, the linear and logit parts of the model separate, making W_{ij} independent of $Z_{ij'}$ for all $j \neq j'$. In the context of adolescent substance use, separability would imply that the presence or absence of use at one occasion has no influence on the amount of use, if any, at other occasions. In our analyses we have found that this condition typically does not hold; random effects from two parts are usually correlated.

4.2 Computational algorithms

Fitting the model (9)–(11) is not trivial. The computations required are similar to those needed for generalized linear models with random effects, sometimes called generalized linear mixed models (GLMMs) (Stiratelli, Laird, and Ware, 1984). EM algorithms, which are straightforward to implement for linear random-effects models, are difficult for GLMMs because the expectations required for the E-step cannot be calculated in closed form. These expectations can be approximated by Taylor linearization or Monte Carlo methods (McCulloch, 1997). Convergence of EM for random-effects models can be painfully slow. We have implemented EM for the model (9)–(11) but found it to be too slow for practical use (Olsen and Schafer, 1998). A variety of Markov chain Monte Carlo (MCMC) procedures have been applied to GLMMs as well (Zeger and Karim, 1991; Clayton, 1996) but these too may suffer from slow convergence, particularly in large samples.

One of the main difficulties with this model is that the likelihood function contains

integrals which cannot be evaluated in closed form. The likelihood is

$$L \propto \prod_{i=1}^{m} \int \exp\{l_{W_i}\} \exp\{l_{Z_i}\}\, p(b_i) db_i,$$

where $l_{W_i} = \sum_{j=1}^{n_i} (W_{ij}\eta_{ij} + \log(1-\pi_{ij}))$ comes from the logistic regression,

$$l_{Z_i} = -\frac{n_i^*}{2}(\log \sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^{n_i^*} (Z_{ij} - (X^*{}_{ij}\gamma + V^*{}_{ij}d_i))^T (Z_{ij} - (X^*{}_{ij}\gamma + V^*{}_{ij}d_i))$$

arises from the linear regression, and $p(b_i) = |\psi|^{-\frac{1}{2}} \exp\{-\frac{1}{2} b_i^T \psi^{-1} b_i\}$ is the normal density applied to the random effects. A variety of methods are available for approximating these integrals, including Gauss-Hermite quadrature (Anderson and Aitkin, 1985) and Laplace approximations (Tierney and Kadane, 1986). Using techniques devised by Raudenbush, Yang and Yosef (1999), we have implemented a sixth-order multivariate Laplace approximation to the loglikelihood which is accurate and fast. We maximize this loglikelihood by an approximate Fisher scoring procedure which we have coded in a Fortran-90 program. Upon convergence, the program provides maximum-likelihood estimates and a joint covariance matrix for β, γ, and the free parameters in ψ. Our program also calculates empirical Bayes estimates for the random effects b_i and their covariances by importance sampling. Detailed descriptions of these methods are provided by Olsen and Schafer (under review).

Although these new algorithms were developed primarily for data analysis and parameter estimation, it would be relatively easy to adapt them for imputation of missing responses. Imputation could be carried out by the following steps.

1. Fit the two-part random-effects model to obtain an estimate $\hat{\theta}$ for $\theta = (\beta, \gamma, \psi)$.

2. Draw a random value θ^* from an approximate posterior distribution centered at $\hat{\theta}$ with covariances obtained from the scoring procedure.

3. Use θ^* to calculate an empirical Bayes mean and covariance matrix for each random effect b_i. Draw a new random effect b_i^* from a normal distribution with this mean and covariance independently for $i = 1, \ldots, m$.

4. For each sample unit $i = 1, \ldots, m$, substitute the simulated random effect b_i^* and simulated parameters θ^* into (9) and (10) to obtain prediction equations for W_{ij} and Z_{ij}. Draw a Bernoulli variate W_{ij}^* and a normal variate Z_{ij}^* for each occasion where the semicontinuous response Y_{ij} is missing.

5. Set $Y_{ij}^* = 0$ if $W_{ij}^* = 0$ and $Y_{ij}^* = g^{-1}(Z_{ij}^*)$ if $W_{ij}^* = 1$.

Repeating these steps M times would produce a set of M multiple imputations for the missing values of Y_{ij}.

4.3 Example: parental monitoring and reported alcohol use

In the Adolescent Alcohol Prevention Trial (AAPT; Hansen and Graham, 1991), students in Los Angeles and Orange Counties were surveyed annually in grades 7–11 on the use of alcohol and other controlled substances, attitudes toward school, personality traits and family characteristics. Using our two-part random-effects model, we explored the relationships between reported recent alcohol use over grades 7–11 and the degree of parental monitoring reported in grade 7. A low level of monitoring is believed to be a potential risk factor for illicit substance use. Among the AAPT subjects, we have found that that this risk factor seems to operate quite differently on the probability of alcohol use versus the amount of alcohol consumption when it occurs.

Our response variable is a composite measure of reported recent alcohol use, with $Y_{ij} = 0$ representing no use or sips for religious purposes only. To reduce skewness, we applied a log transformation to the nonzero responses, taking $Z_{ij} = \log Y_{ij}$ if $Y_{ij} > 0$. The measure of parental monitoring is a standardized composite of three items recorded in grade 7: (1) When you go out with your friends, how often do your parents tell you what time to be home? (2) How often do your parents refuse to let you go places and do things with other people your age? (3) How often do your parents ask where you are going when you leave

Table 1: Estimated coefficients and standard errors from two-part random-effects model for reported recent alcohol use

	$\hat{\beta}$	SE($\hat{\beta}$)	$\hat{\gamma}$	SE($\hat{\gamma}$)
intercept	-1.941	0.101	-0.282	0.050
sex	-0.353	0.145	0.042	0.061
monitoring	-0.415	0.136	-0.057	0.051
time	0.351	0.029	0.171	0.014
monitoring × sex	0.461	0.206	-0.045	0.081
time × sex	0.063	0.044	0.014	0.022
time × monitoring	0.161	0.042	-0.010	0.020
time × monitoring × sex	-0.220	0.065	0.037	0.030

the house? Higher values for this composite variable correspond to increasing levels of supervision.

In our model, the columns of X_i and X_i^* are identical and include an intercept, main effects for sex, time, and parental monitoring, all possible two-way interactions among them, and the three-way interaction. Time is coded from 0=grade 7 to 4=grade 11 and gender is coded as 0=female and 1=male. We first attempted to fit models which allowed the slopes to vary by subject, but in each case the estimated variances of the slopes became zero. Our final model allows only the intercepts for each part to vary by subject, so that V_i and V_i^* are simply vectors of 1's. The scoring procedure converged in 38 iterations to a maximum relative parameter change of 0.001. Execution on a Pentium II 400 Mhz workstation took 52 seconds. Maximum-likelihood estimates and standard errors for β and γ are shown in Table 1. The variance-component estimates $\hat{\psi}_{cc} = 3.49$, $\hat{\psi}_{cd} = 0.648$, and $\hat{\psi}_{dd} = 0.214$ reveal substantial individual variation in both the probability of use and in the expected amount of use if any.

To clarify the role of parental monitoring, we calculated the estimated effect of a one standard-unit decrease in monitoring for boys and girls in each grade. For the logit model, the effect is expressed as an odds ratio; for the linear model the effect is a mean difference. These two sets of effects are plotted in Figure 6. Consider first the plot of the odds ratio.

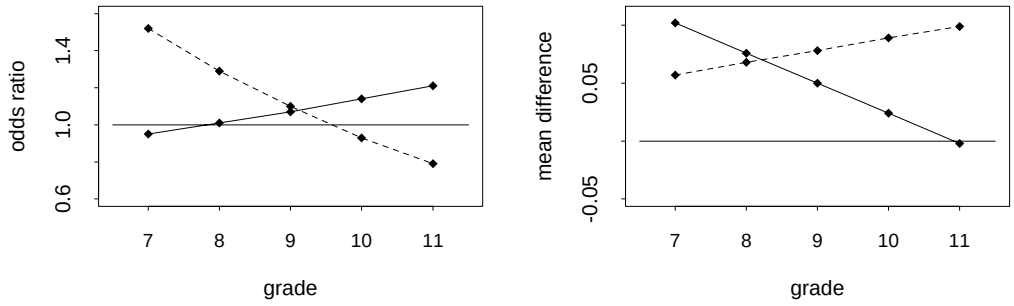

Figure 6: Estimated effect of reduced parental monitoring on the odds of reported recent alcohol use and on the mean level of reported use, if any, for girls (- - -) and boys (—)

For girls, low monitoring in grade 7 appears to substantially increase the odds of alcohol use in grade 7, but the effect diminishes rapidly over time; by grade 11 the girls with less monitoring have *lower* odds of use than those who were monitored more closely. For boys, however, the effects are nearly opposite: low monitoring in grade 7 has little effect at that time, but by grade 11 the less monitored boys are using alcohol at substantially higher rates than the highly monitored ones. The second plot in Figure 6, which shows the effect of low monitoring on reported amounts of alcohol consumption, tells quite a different story. In grade 7, low levels of monitoring increase the amount of reported use when it occurs for both boys and girls. But among girls this effect tends to increase over time, whereas for boys it decreases and vanishes by grade 11.

5 Multivariate semicontinuous data

5.1 Iterative univariate imputation

We now consider more general survey problems where multiple categorical, continuous and semicontinuous variables have been recorded with occasional missing values. Our goal is no longer to analyze or describe the relationships among these variables in a scientifically meaningful way, but to devise a procedure for imputing the missing values which preserves

important features of the marginal and joint distributions. Proposing a joint model for multivariate categorical, continuous and semicontinuous data is not difficult, but fitting the model can be challenging, especially when the number of response variables is large. Explicitly multivariate models will be discussed momentarily. First, we mention some approximate methods based on iterative univariate imputation.

Iterative univariate imputation was first implemented for the Survey of Consumer Finances (Kennickell, 1991; Kennickell and McManus, 1994). The method is motivated by a popular MCMC procedure known as Gibbs sampling (Gelfand and Smith, 1990). Gibbs sampling simulates the joint distribution of random variables $Y = (Y_1, \ldots, Y_r)$ by iteratively drawing from the conditional distribution of each Y_j given

$$Y_{(-j)} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_r). \tag{12}$$

for $j = 1, \ldots, r$, substituting the most recently simulated values of $Y_1, \ldots, Y_r$ in (12) at each draw. Simulating Y_j given $Y_{(-j)}$ for $j = 1, \ldots, r$ comprises one cycle of the Gibbs sampler and produces one simulated value for Y. Repeating the process many times creates a sequence of Y-values from a Markov chain whose stationary law is the joint distribution of $Y = (Y_1, \ldots, Y_r)$. For more information on Gibbs samplers, see Casella and George (1992) and Gilks, Richardson and Spieglhalter (1996).

Kennickell (1991) adapts this idea to imputation of survey variables in the following way. In many surveys, it may be difficult to propose a sensible joint distribution for all variables of interest. But given the variety of procedures available for regression modeling of continuous and categorical responses, it may be possible to specify a plausible regression for predicting one variable given the others. Let $Y_j = (y_{1j}, \ldots, y_{nj})$ represent the vector of values for survey variable j for sample units $i = 1, \ldots, n$. Let obs_j and mis_j denote the sets of sample units for which variable j is observed and missing, respectively. If complete data for all other variables $Y_{(-j)} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_r)$ were available, then Bayesian imputations for $\{y_{ij}, i \in mis_j\}$ given $Y_{(-j)}$ could be created as follows.

1. Regress variable j on all other variables plausibly related to it using an appropriate (e.g. linear or generalized linear) model, fitting the model to sample units $i \in obs_j$. This produces an estimate $\hat{\theta}_j$ of the unknown regression parameters θ_j.

2. Draw a simulated value θ_j^* from an approximate posterior distribution for θ_j. The posterior may be approximated, for example, by a multivariate normal distribution centered at the maximum-likelihood estimate $\hat{\theta}_j$ with covariances determined by the second derivatives of the loglikelihood function.

3. Substitute θ_j^* into the regression function and draw simulated values for $\{y_{ij}, i \in mis_j\}$ from the implied predictive distribution given $Y_{(-j)}$.

Now because $Y_{(-j)}$ itself contains missing values, steps 1-3 cannot be carried out unless these values have been imputed. This suggests, by analogy to a Gibbs sampler, that one could jointly impute $Y_1, \ldots, Y_r$ by an iterative scheme. To begin, we fill in all missing values in $Y_1, \ldots, Y_r$ by a simple method, for example, by univariate hot-deck imputation. Then we perform steps 1–3 above for each Y_j, $j = 1, \ldots, r$ in turn, using the most recently imputed data for $Y_{(-j)}$ at each step. As we repeat this imputation cycle over and over, the imputed data gradually acquire the inter-variable relationships described by the regression models. After a sufficient number of cycles, the system may stabilize to a set of joint relationships among $Y_1, \ldots, Y_r$.

Iterative univariate imputation is quite flexible and can readily accommodate a large number of continuous and categorical variables. Semicontinuous variables are easily handled by two-part regression, using the imputation procedure described at the end of Section 3.1. General-purpose SAS routines for iterative univariate imputation have been developed by Raghunathan and Lepkowski (1999).

Despite the attractive features of this method, its statistical properties are not well understood because the joint distribution for $Y_1, \ldots, Y_r$ may not actually exist. In a true

Gibbs sampler, one begins with a possibly intractable but well-defined joint distribution for $Y_1, \ldots, Y_r$ and samples from the actual conditional distributions for Y_j given $Y_{(-j)}$, $j = 1, \ldots, r$ implied by this joint distribution. In iterative univariate imputation, however, one proposes a reasonable conditional distribution for each Y_j given $Y_{(-j)}$ in the hope that these conditional distributions define a suitable joint model. It is easy to construct examples of conditional distributions which do not define an actual joint distribution (e.g. Casella and George, 1992). In such cases, iteratively sampling from these 'conditionals' creates a Markov chain that never actually converges; the probability law describing the state of the system does not stabilize but continues to change from cycle to cycle. The practical implications of this phenomenon in iterative univariate imputation are unknown.

5.2 Extending the general location model

Extensions of the general location model to handle a single semicontinuous response were covered in Section 3.2. In principle, the general location model can handle any number of semicontinuous variables at once. Each semicontinous variable Y_j can be recoded into a binary indicator W_j and a continuous variable Z_j which is regarded as 'missing' whenever $Y_j = 0$. A relationship between W_j and Z_j would have little meaning and could not be estimated, because Z_j is never seen when W_j takes one of its two values. These inestimable relationships could be removed from the model as described in Section 3.2 by changing (8) from a standard regression model to a seemingly unrelated regression.

Seemingly unrelated regression models can be fit by an iterative, two-step method which alternates between (a) estimating the coefficients β by generalized least squares (GLS) under an assumed value for the residual covariance matrix Σ, and (b) re-estimating the residual covariances from the matrix of residuals implied by the new estimate for β (Zellner, 1962). It is theoretically possible to incorporate this two-step procedure into the model-fitting and imputation algorithms described by Schafer (1997, ch. 9), generalizing them to handle any

number of continuous, categorical, and semicontinuous variables. In practice, however, we have found three computational issues which make an implementation troublesome. The first difficulty is that the matrix which must be inverted for GLS estimation tends to be large. A standard regression model requires inversion of a symmetric $s \times s$ matrix where s is the number of regressors. Seemingly unrelated regression, however, inverts a matrix of size $S \times S$, where $S = \sum_{j=1}^{q} s_j$ and s_j is the number of regressors for response variable Z_j, $j = 1, \ldots, q$. As a result, the GLS procedure becomes exceedingly expensive as the number of variables grows.

A second issue associated with this method is that rates of 'missing information' tend to grow rapidly as semicontinuous variables are added to the model. Because the continuous portion Z_j of a semicontinuous variable Y_j is treated as missing whenever $Y_j = 0$, the means, variances, and especially correlations among the Z_j are difficult to estimate when the point masses $P(Y_j = 0)$ are fairly large. In theory, these high rates of missing information are not necessarily a problem because the 'missing' values of Z_j are actually irrelevant to the statistical model for Y_j; they are being treated as missing purely for computational convenience. Yet, the resulting iterative estimation and simulation procedures may converge slowly because convergence behavior is typically governed by the highest rate of missing information in the system (e.g. Schafer, 1997, ch. 3).

A third problem that makes the general location model difficult to apply is that the number of cells in the contingency table explodes as the number of categorical variables increases. Each semicontinuous Y_j added to the system contributes a binary indicator W_j which causes the number of cells to double. Iterative algorithms which operate on cell counts grind to a halt as more categorical variables are added. Even with advanced computers, fitting a general location model with 20 or more categorical and semicontinuous variables may not be feasible for some time.

5.3 Imputation under a non-identified normal model

Whenever a semicontinuous variable is recoded as a binary indicator and a partially missing continuous variable, a relationship between these two variables cannot be directly observed due to the missing-data pattern. Yet, if our goal is to generate plausible imputations for the original semicontinuous variable, then such a relationship is superfluous. Let Y be a semicontinuous response which we recode as $W = 1$ and $Z = g(Y)$ if $Y \neq 0$, $W = 0$ and Z missing if $Y = 0$. The only aspects of the joint distribution for (W, Z) relevant to Y are the marginal distribution for W and the conditional distribution for Z given $W = 1$. As shown by (7), any assumptions made about the conditional distribution of Z given $W = 0$—and by extension, about the relationship between W and Z—are unverifiable and quite irrelevant.

In a multivariate setting, the inestimability of the W-Z relationship for each semicontinuous Y leads to complicated fitting procedures if we specify a joint model in which these relationships are nullified. But what happens if we specify a model in which these relationships are allowed but remain inestimable? EM algorithms are quite stable and tend to converge reliably to maximum-likelihood estimates even when the estimates are not unique. MCMC procedures for posterior simulation may fail to converge if an improper prior distribution is applied, because a proper posterior distribution may not exist. Yet the non-existence of a joint posterior for all parameters in the model does not necessarily imply non-existence of a marginal posterior for relevant parameters of interest. An MCMC algorithm applied with a vague prior might fail to converge in a global sense, yet behave well with respect to the parameters of interest.

Based upon this observation, it may be quite reasonable to describe W-Z pairs by a non-identified multivariate model for the purpose of imputing missing values for Y. If the number of variables is not too large, then a general location model with a standard regression (8) is a natural choice. As the number of variables grows, general location models

become infeasible due to the large number of cells in the contingency table. In practice, we have found that reasonable imputations can often be obtained from algorithms based on a multivariate normal model. The binary indicators W and the continuous variables Z are entered into the model together and imputed as if they were jointly normal, and the continuously distributed imputes for W are rounded to zero or one. As a final step, imputations for Y are obtained by setting $Y = 0$ if $W = 0$ and $Y = g^{-1}(Z)$ if $W = 1$. Incomplete-data modeling and imputation procedures for the multivariate normal model are well developed and a variety of free software is available. A stand-alone Windows program called NORM combines model-fitting and imputation procedures with an easy-to-use Windows interface (Schafer, 1999). The computational algorithms in NORM are also distributed as a library of functions for S-PLUS (Schafer, 1996c).

Treating Bernoulli variates as normal for purposes of imputation seems ad hoc, but the method has been shown to work well in several simulation studies (Schafer, 1997). More rigorous procedures are possible if we assume that W is a recode of an underlying normal variate R, where $W = 0$ if $R \leq 0$ and $W = 1$ if $R > 0$, and model the (R, Z) pairs as jointly normal. The latter approach is equivalent to specifying a probit regression for W given Z. But because R is never directly observed, the computational algorithms needed to handle the probit version—the EM algorithm in particular—are much more complicated than those that simply treat W as normal.

When rounding the continuous imputed values for W to zero or one, a cutoff value of $c = 0.5$ works well when the overall proportion of zeros is not too extreme (10–90%). Better results may be obtained with the adaptive cutoff $c(\bar{w}) = \bar{w} - \Phi^{-1}(\bar{w})\sqrt{\bar{w}(1-\bar{w})}$, where $\bar{w}$ is the mean of the W values prior to rounding and Φ^{-1} denotes the standard normal quantile function (e.g. $\Phi^{-1}(.975) = 1.96$). A plot of $c(\bar{w})$ versus $\bar{w}$, shown in Figure 7, reveals that $c(\bar{w}) \approx 0.5$ for values of $\bar{w}$ between 0.1 and 0.9.

To demonstrate the method, we applied the normal model footwear expenditures data

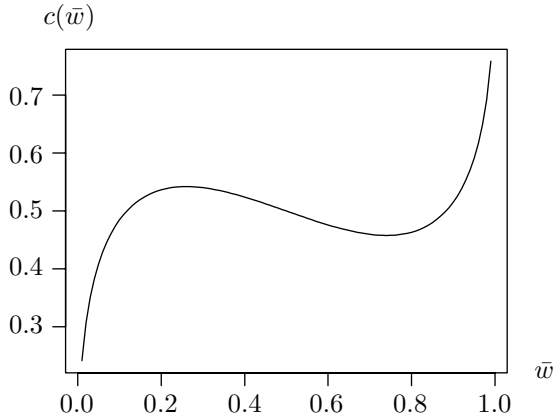

Figure 7: Adaptive cutoff $c(\bar{w})$ for rounding imputed values as function of $\bar{w}$, the observed mean of W prior to rounding.

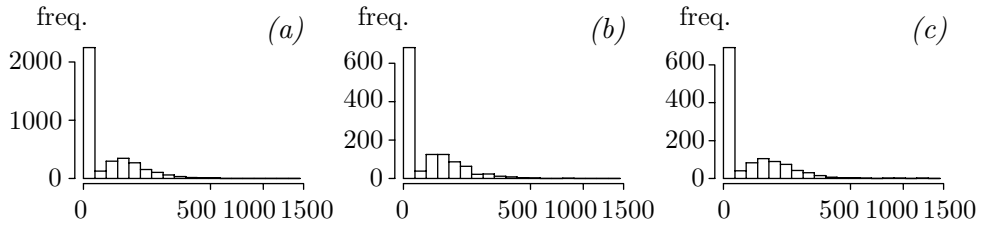

Figure 8: Dollar expenditures on footwear reported in the Consumer Expenditures Survey during the first quarter of 1996: (a) observed values, (b) first imputation, (c) second imputation.

shown in Figure 1 (a), recoding the variable Y as $W = 1$ and $Z = \log Y$ if $Y \neq 0$, $W = 0$ and Z missing if $Y = 0$. The log transformation is a natural choice for an expenditure variable because, if the data are imputed on a log scale and transformed back to the original scale, the imputed expenditure values will never be negative. We randomly deleted 25% of the values for Y to induce additional missing values for W and Z. NORM's EM algorithm converged very quickly to non-unique maximum-likelihood estimates for the normal parameters. We then ran data augmentation, a two-step Gibbs sampler for parameter simulation and multiple imputation, stopping every 1,000 cycles to impute the missing values. Histograms of the observed expenditures and two sets of imputed data are shown in Figure 8.

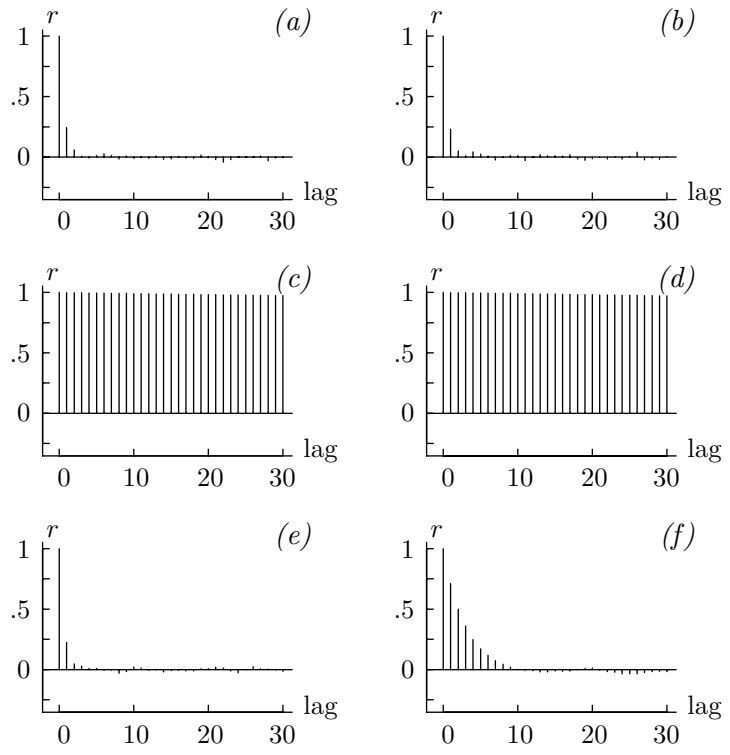

Figure 9: Sample autocorrelation functions for parameters of bivariate normal distribution applied to W and Z from 5,000 cycles of data augmentation: (a) μ_W, (b) σ^2_W, (c) μ_Z, (d) ρ, (e) $E(Z \mid W = 1)$ and (f) $V(Z \mid W)$.

NORM's data augmentation algorithm behaves erratically under the default noninformative prior distribution because certain aspects of the bivariate normal model are not identified. In routine applications of data augmentation, one hopes to see autocorrelations in the simulated parameters die down to zero after a reasonable number of cycles; if the lag-k autocorrelations for all parameters are essentially zero, then imputations generated at every k cycles will be approximately independent. In this model, however, the autocorrelations for some parameters do not die down no matter how many cycles are taken. Sample autocorrelation functions (ACFs) for some parameters estimated from 5,000 cycles of data augmentation are displayed in Figure 9. Plots (a) and (b) show ACFs for $\mu_W = E(W)$ and $\sigma^2_W = V(W)$, which diminish quickly. But the ACFs for μ_Z and $\rho = \text{Corr}(W, Z)$ shown in plots (c)–(d) do not die down, because these parameters are influenced by a W-Z relationship that cannot be estimated. Time series plots for μ_Z and ρ (not dis-

played) show the parameters drifting aimlessly through space. Parameters pertaining to the conditional distribution of Z given $W = 1$, however, are well behaved. ACFs for $E(Z \mid W = 1) = \mu_Z + (1 - \mu_W)\text{Cov}(W, Z)/\sigma_W^2$ and $V(Z \mid W) = (1 - \rho^2)V(Z)$, are shown in plots (e)–(f). These ACFs drop quickly, suggesting that this model does indeed give plausible imputations for Y. In light of plots (a), (b), (e), and (f), it appears that taking 1,000 cycles of data augmentation between imputations was not at all necessary, and quality imputations could have obtained with as few as 10 cycles.

In this application, the lack of identification in the bivariate model for (W, Z) does not impair the implied predictive distribution for Y. When running data augmentation with an unidentified model, there is a practical danger that one or more non-identified parameters may drift to extremely large values and produce numeric overflow. To prevent this from happening, we may introduce a small amount of prior information to stabilize the covariances. As an alternative to the default noninformative prior, NORM allows the user to specify a ridge prior which smooths the estimated correlations toward an identity matrix (Schafer, 1997, ch. 5). The degree of smoothing is controlled by a single hyperparameter which can be interpreted as a number of prior joint observations of (W, Z). Setting this hyperparameter to a small positive value—say, 1% of the sample size or less—will usually be sufficient to stabilize the computations and prevent overflow. Larger values for the hyperparameter may introduce bias into the identified parts of the model and should be avoided.

5.4 A multivariate simulation study

The procedure described above for one semicontinuous variable extends immediately to any number of variables. To see how well the method performs in a realistic multivariate setting, we conducted a simulation study with data drawn from the Consumer Expenditures (CE) Survey from the first quarter of 1996. CE participants report income and dollar

Table 2: Six expenditure variables related to household furnishings and three related covariates from the Consumer Expenditures Survey

TEXTIL	Household textiles
FURNTR	Furniture
FLRCVR	Floor coverings
MAJAPP	Major appliances
SMLAPP	Small appliances and miscellaneous housewares
MISCEQ	Miscellaneous household equipment
FAMSIZE	Number of persons in the consumer unit
CUTENURE	Home ownership or rental status
FINCATAX	Consumer unit income after taxes

amounts spent on all types of goods and services; reported expenditures are then aggregated into broader categories such as food and clothing. For this simulation, we focused on six interrelated expenditure variables which comprise household furnishings and equipment. These six variables are listed in Table 2. All six are semicontinuous; unweighted estimates of the proportions of zero values range from 0.58 for MISCEQ to 0.96 for FLRCVR. We also included three characteristics of the family or consumer unit that may be strongly related to the level of expenditures: number of persons in the consumer unit, home ownership or rental status, and income after taxes.

Public-use files from the Bureau of Labor Statistics provided values of these variables for $N = 4,876$ consumer units. Treating this group as an artificial population, we drew simple random samples of $n = 500$ without replacement and randomly imposed missing values on the expenditures variables at an average 25% rate. Multiple imputations were created under a non-identified multivariate normal model as described in Section 5.3, applying log transformations to the positive values of each expenditure. The imputation model contained fifteen variables: two for each type of expenditure, log-transformed versions of FAMSIZE and FINCATAX, and a binary recode of CUTENURE indicating ownership or non-ownership. Imputations were created in the following manner. First, initial parameter estimates were obtained by running EM until convergence up to a maximum of 100 itera-

tions. Using the EM estimates as starting values, we ran 500 cycles of data augmentation under a ridge prior with the hyperparameter set to 2.0, stopping after every 100 cycles to produce $M = 5$ multiple imputations. Imputed values were rounded using the adaptive cutoff rule described in Section 5.3. Finally, using the method for scalar estimands given by Rubin (1987), point and interval estimates were calculated for various quantities of interest—means, proportions, odds ratios, correlation and regression coefficients—and the estimates and intervals were compared to the corresponding population values. No corrections for finite-population sampling were used. The entire sampling, imputation and estimation procedure was repeated 1,000 times.

For evaluating the performance of the imputation procedure, we mainly considered bias in point estimates and actual coverage of nominal 95% intervals. One point of concern is whether modeling binary indicators as normal and rounding off the imputed values distorts the proportions of zeros in the expenditure variables. Another concern is whether the model is able to capture various interrelationships among the expenditure variables such as those depicted in Figure 2. Model failure could bias point estimates and produce coverages lower than the nominal rates. But poor performance may also result if the estimation and interval procedures are themselves biased apart from the missing-data aspect. Estimates for nonlinear quantitites such as odds ratios and correlation coefficients are approximately unbiased only if the sample size is sufficiently large. Even a simple confidence interval for a mean can perform poorly when sampling from a population with extreme outliers. To help assess the performance of the imputation method apart from the performance of the complete-data inference procedure, we also calculated estimates and intervals from each sample of $n = 500$ before missing values were imposed.

Results from the simulation are summarized in Table 3. For each estimand, this table shows the actual population value, average estimate, bias of the estimate, and interval coverage for the estimation procedure without missing data and with multiple imputation.

Table 3: Simulated performance of complete-data and multiple-imputation procedures for various estimands from the Consumer Expenditures Survey: population value, average estimate, relative bias as a percentage of the average standard error, and coverage of the nominal 95% interval

		Complete data			*With imputation*		
Estimand	pop.	est.	bias	cvg.	est.	bias	cvg.
(a) Percent without expenditure							
TEXTIL	75.9	75.9	0.0	961	75.8	−5.3	962
FURNTR	88.1	88.1	1.0	950	88.0	−7.4	950
FLRCVR	95.9	95.8	−0.9	963	95.8	−6.3	940
MAJAPP	90.8	90.8	2.5	963	90.7	−6.7	942
SMLAPP	75.6	75.6	−2.8	968	75.5	−6.7	963
MISCEQ	54.7	54.7	−1.4	962	54.9	6.6	949
(b) Mean expenditure							
TEXTIL	24.0	24.1	2.6	893	25.0	16.7	940
FURNTR	84.6	84.4	−0.6	918	108.0	38.7	957
FLRCVR	11.7	11.7	0.4	819	76.7	32.2	877
MAJAPP	39.5	39.5	−0.5	937	46.9	35.6	969
SMLAPP	19.5	19.6	2.4	923	20.4	21.3	951
MISCEQ	136.0	136.1	0.1	906	138.8	8.1	919
(c) Log-odds ratio							
$\texttt{TEXTIL}_{>0} \times \texttt{MISCEQ}_{>0}$	1.47	1.46	−2.5	955	1.30	−57.7	955
$\texttt{TEXTIL}_{>0} \times \texttt{FLRCVR}_{>0}$	1.53	1.54	3.6	969	1.31	−36.3	968
$\texttt{FURNTR}_{>0} \times \texttt{FLRCVR}_{>0}$	0.87	0.85	−2.3	977	0.70	−23.0	978
$\texttt{FURNTR}_{>0} \times \texttt{MAJAPP}_{>0}$	1.11	1.10	−2.7	967	0.96	−30.7	964
(d) Regression coefficient							
$\texttt{MISCEQ}_{+}$ on $\texttt{SMLAPP}_{>0}$	45.3	44.7	−0.5	956	73.0	17.0	983
$\texttt{MISCEQ}_{+}$ on $\texttt{TEXTIL}_{>0}$	65.5	66.7	1.1	944	87.9	13.6	977
$\texttt{SMLAPP}_{+}$ on $\texttt{MISCEQ}_{>0}$	26.2	26.6	1.2	975	25.3	−2.7	990
$\texttt{SMLAPP}_{+}$ on $\texttt{TEXTIL}_{>0}$	16.9	17.4	2.3	961	15.4	−4.8	982
(a) Correlation coefficient							
$\texttt{TEXTIL}_{+}$ with $\texttt{SMLAPP}_{+}$	0.20	0.24	29.0	868	0.21	8.7	971
$\texttt{TEXTIL}_{+}$ with $\texttt{MISCEQ}_{+}$	0.11	0.16	39.2	897	0.17	34.6	976
$\texttt{SMLAPP}_{+}$ with $\texttt{MISCEQ}_{+}$	0.04	0.05	12.2	916	0.09	13.2	973

Bias is expressed as percentage of the average standard error; for example, a value of 50% indicates a bias equivalent to about one-eighth of the width of a nominal 95% interval.

Section (a) of Table 3 displays results for the estimated percentage of consumer units reporting no expenditure in a given category. The complete-data procedure performs well; it is unbiased and has simulated coverage rates of 95% or more. The multiple-imputation procedure also performs very well; the estimates show very slight biases equivalent to 5–7% of one standard error, and coverage rates are close to 95%. It appears that imputing binary indicators under the normal model and rounding them by our adaptive cutoff rule produces very little distortion.

Part (b) of Table 3 pertains to mean expenditures. The complete-data procedure is again unbiased, but its coverage rates for some variables are poor. For these variables, the population values are highly skewed with a few extreme outliers which exert a great deal of influence over the mean when they are sampled, causing the complete-data intervals to behave erratically. The multiple-imputation estimates show some moderate biases up to about 40% of one standard error, equivalent to one-tenth the width of a nominal 95% interval; however, the coverage of the multiple-imputation intervals is actually better than that of the complete-data intervals. The biases in the estimated means may be an artifact of the log transformation used in the imputation procedure. These biases could perhaps be reduced by imputing without transformation, but this might distort other aspects of distributional shape.

Parts (c)–(e) of Table 3 show results for various aspects of relationships among the semi-continuous expenditures variables. In these sections, a subscript '0 >' indicates a binary indicator for positive expenditures, and a subscript '+' refers to the subset of expenditure values that are positive. The multiple-imputation estimates for log-odds ratios are moderately biased toward zero, but the actual coverage rates of the interval estimates equal or exceed 95%. Complete-data and multiple-imputation procedures for regression coeffi-

cients perform well. Finally, the complete-data procedures for correlation coefficients have difficulty, partly due to the small sample sizes (relatively few respondents report non-zero expenditures in both categories) and partly due to the skewness in the population distributions. The multiple-imputation procedures for correlations fare better, especially with regard to interval coverage rates.

In summary, imputing expenditures based on the non-identified normal model appears performs quite well. Further refinements to improve performance are certainly possible, but the simplicity of this approach and the ease of implementation with existing software make it an attractive method for moderately sized data sets. The Windows version of NORM has been used routinely in data sets with more than 100 variables. With a very large number of variables (say, 200 or more) one may need to impose additional structure on the covariance matrix to reduce the number of unknown parameters.

6 Discussion

In this article we have reviewed a variety of methods for modeling and imputing semicontinuous variables singly and jointly. In many cases, we are able to impute these variables with models that explicitly account for their two-part nature. More work is needed, however, before these methods can be applied routinely by practitioners of large-scale national surveys. In many survey applications, it will be necessary to take into account important aspects of the complex sample design. Models which allow intracluster correlations among units may be necessary to impute a cluster survey. Multivariate extensions of linear random-effects models are now being developed along with the necessary algorithms and software (Schafer, 1998). These models may also be used for panel surveys, where interrelated variables are measured for subjects at multiple points in time. Current imputation procedures for panel surveys often do not take longitudinal structure into account, but impute each wave separately as if it were a cross-sectional dataset. The ability to model individual behavior

longitudinally and pool information across waves may substantially improve the precision of many survey estimates, especially if present behavior is highly correlated with the past.

Future techniques should address not only the problem of missing values but other types of data censoring, collapsing and coarsening. In certain contexts, it is quite common for respondents to have only partial information about a quantity of interest. For example, a subject may know that he spent more than $100 on auto fuel but may be unable to give a more precise amount. Rather than pressuring the subject to 'make up' a more precise value, it may be better to simply record the lower bound of $100 with an indication that the response has been right-censored. More flexible data-collection protocols will require more complicated imputation procedures. In the CE Survey, subjects are sometimes able to provide a total dollar amount spent on a broad class of goods or services but cannot break it down further into the narrow categories desired by the data collector. In these cases, an imputation method must allocate the fixed total amount to the relevant subclasses. One can envision a multivariate model-based procedure capable of imputing a group of interrelated semicontinuous variables subject to an equality constraint placed on their sum. Implementing algorithms for this type of data may be especially challenging when the data are modeled after transformation, because applying logarithmic or other transformations to variables will make such constraints nonlinear.

References

Amemiya, T. (1984) Tobit models: a survey. *Journal of Econometrics*, 24, 3–61.

Anderson, D.A. and Aitkin, M. (1985) Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society*, Series B, 47, 204–210.

Casella, G. and George, E.I. (1992) Explaining the Gibbs sampler. *The American Statistician*, 46, 167–174.

Clayton, D.G. (1996) Generalized linear mixed models. *Markov Chain Monte Carlo in*

Practice (eds W.R. Gilks, S. Richardson and D.J. Spiegelhalter), 275–301. Chapman & Hall, London.

Duan, N., Manning, W. G., Morris, C. N. and Newhouse, J. P. (1983) A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics*, 1, 115-126.

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398–409.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., eds (1996) *Markov-Chain Monte Carlo in Practice*. Chapman & Hall, London.

Hansen, W.B. & Graham, J.W. (1991). Preventing alcohol, marijuana, and cigarette use among adolescents: Peer pressure resistance training versus establishing conservative norms. *Preventive Medicine*, 20, 414–430.

Hartley, H.O. and Rao, J.N.K. (1967) Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54, 93–108.

Heckman, J. (1974) Shadow prices, market wages, and labor supply. *Econometrica*, 42, 679-674.

Heckman, J. (1976) The common structure of statistical models of truncation, sample selection, and limited dependent variables, and a sample estimator for such models. *The Annals of Economic and Social Measurement*, 5, 475-592.

Jennrich, R.I. and Schluchter, M.D. (1986) Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 38, 967–974.

Kennickell, A.B. (1991) Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1–10.

Kennickell, A.B. and McManus, D.A. (1994) Multiple imputation of the 1983 and 1989 waves of the SCF. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 523–528.

Laird, N.M., Lange, N. and Stram, D. (1987) Maximum likelihood computations with

repeated measures: application of the EM algorithm. *Journal of the American Statistical Association*, **82**, 97–105.

Laird, N.M. and Ware, J.H. (1982) Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.

Lindstrom, M. J. and Bates, D.M. (1988) Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83, 1014–1022.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996). *SAS System for Mixed Models*. SAS Institute, Cary, NC.

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

Manning, W., Morris, C.N., and Newhouse, J.P. (1981) A two-part model of the demand for medical care: Preliminary results from the health insurance experiment. *Health, Economics, and Health Economics* (eds. J. van der Gaag and M. Perlman), 103-124, North-Holland, Amsterdam.

MathSoft, Inc. (1997) *S-PLUS User's Guide*, Data Analysis Products Division, MathSoft, Seattle, WA.

McCulloch, C.E. (1997) Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92, 162-170.

NHTSA (1998) Research Note: Multiple imputation of missing blood alcohol content (BAC) values in FARS. Washington, DC: National Highway Traffic Safety Administration, U.S. Department of Transportation.

Olsen, M.K. and Schafer, J.L. (1998) Parmater estimates for semicontinuous longitudinal data using an EM algorithm. *Technical Report* # 98-31, The Center for the Study of Prevention through Innovative Methodology, The Pennsylvania State University.

Olsen, M.K. and Schafer, J.L. (under review) A two-part random-effects model for semicontinuous longitudinal data. Submitted to *Journal of the American Statistical Association*.

Raghunathan, T.E. and Lepkowski, J. (1999) Multiple imputation for a larg scale national

survey. Presented at Joint Statistical Meetings, August 1999, Batimore.

Raudenbush, S.W., Yang, M., and Yosef, M. (1999) Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, in press.

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

Schafer, J.L. (1996a) *CAT: Multiple imputation for multivariate categorical data*, software library for S-PLUS. Written in S-PLUS and Fortran-77. Available at `http://www.stat.psu.edu/~jls/`.

Schafer, J.L. (1996b) *MIX: Multiple imputation for mixed continuous and categorical data*, software library for S-PLUS. Written in S-PLUS and Fortran-77. Available at `http://www.stat.psu.edu/~jls/`.

Schafer, J.L. (1996c) *NORM: Multiple imputation for continuous data based on a normal model*, software library for S-PLUS. Written in S-PLUS and Fortran-77. Available at `http://www.stat.psu.edu/~jls/`.

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

Schafer, J.L. (1998) *PAN: Multiple imputation for longitudinal and clustered data under a multivariate linear mixed model*, software library for S-PLUS. Written in S-PLUS and Fortran-77. Available at `http://www.stat.psu.edu/~jls/`.

Schafer, J.L. (1999) *NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2.* Software for Windows 95/98/NT, available from `http://www.stat.psu.edu/~jls/misoftwa.html`.

Stiratelli, R., Laird, N. and Ware, J.H. (1984) Random-effects models for serial observations with binary response. *Biometrics*, 40, 961-971.

Tierney, L. and Kadane, J.B. (1986) Accurate approximations for posterior moments and densities. *Journal of the American Statistical Association*, 81, 82-86.

Tobin, J. (1958) Estimation of Relationships for Limited Dependent Variables. *Econometrica*, 26, 24-36.

Zeger, S.L. and Karim, M.R. (1991) Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association*, 86, 79–86.

Zeger, S.L., Liang, K.Y. and Albert, P.S. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44, 1049–1060.