USING DATA AUGMENTATION TO CORRECT FOR NONIGNORABLE NONRESPONSE WHEN SURROGATE DATA ARE AVAILABLE: AN APPLICATION TO THE DISTRIBUTION OF HOURLY PAY

GABRIELE BEISSEL-DURRANT, CHRIS SKINNER

ABSTRACT

This paper develops a data augmentation method to estimate the distribution function of a variable, which is partially observed, under a nonignorable missing data mechanism, and where surrogate data are available. An application to the estimation of hourly pay distributions using UK Labour Force Survey (LFS) data provides the main motivation.

In addition to considering a standard parametric data augmentation method, we consider the use of hot deck imputation methods as part of the data augmentation procedure to improve the robustness of the method. The proposed method is compared with standard methods based upon an ignorable missing data mechanism, both in a simulation study and in the LFS application. The focus is on reducing bias in point estimation, but variance estimation using multiple imputation is also considered briefly.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M04/10

Using Data Augmentation to Correct for Nonignorable Nonresponse when Surrogate Data are Available: An Application to the Distribution of Hourly Pay

Gabriele Beissel-Durrant and Chris Skinner

University of Southampton

Summary. This paper develops a data augmentation method to estimate the distribution function of a variable, which is partially observed, under a nonignorable missing data mechanism, and where surrogate data are available. An application to the estimation of hourly pay distributions using UK Labour Force Survey (LFS) data provides the main motivation. In addition to considering a standard parametric data augmentation method, we consider the use of hot deck imputation methods as part of the data augmentation procedure to improve the robustness of the method. The proposed method is compared with standard methods based upon an ignorable missing data mechanism, both in a simulation study and in the LFS application. The focus is on reducing bias in point estimation, but variance estimation using multiple imputation is also considered briefly.

Key Words: distribution function estimation; imputation; measurement error; missing data; multiple imputation; rejection sampling.

1. Introduction

Hourly pay can be a difficult variable to measure in household surveys. In the UK Labour Force Survey (LFS) this variable is measured in two ways. First, values of the variable are derived from responses to questions about earnings and hours worked. Second, employees who report that they are paid by the hour are asked directly about their hourly rate of pay. We refer to the two variables as the *derived variable* and the *direct variable*. The characteristics of these variables are discussed in Skinner, Stuttard, Beissel-Durrant and Jenkins (2002). In summary, the derived variable appears to be subject to measurement error, substantial enough to lead to serious bias in the estimation of the distribution of hourly pay, but is subject to almost no item nonresponse; the direct variable is much more accurately measured, but is missing for 50-60% of employees. The problem is how best to use the data on both variables to estimate the distribution of hourly pay. As in Skinner et al. (2002), we shall assume here that the direct variable is subject to no measurement error so the key issue is the nonresponse on the direct variable.

The problem may be formulated as the following general missing data problem. Let y_i be a variable of interest, recorded only for a subset of units i in a sample. Let r_i be the binary variable indicating whether y_i is observed ($r_i = 1$) or not ($r_i = 0$). Let x_i be a variable, which measures y_i with error, but is observed for all units in the sample. The problem is how to use these data to make inference about aspects of the distribution of y_i in the population. A critical issue is the nature of any assumptions about the missing data mechanism. A standard assumption is that the data are *missing at random* (MAR), given the values also of additional (completely) observed variables in a vector w_i (Little and Rubin, 2002). This is the assumption underlying methods developed by Skinner et al. (2002) and used by the UK Office for National Statistics for estimating the distribution of hourly pay from the LFS. If we view (y_i, x_i, w_i, r_i) as a random vector then the MAR assumption is that r_i is conditionally independent of y_i given (x_i, w_i).

If we emphasise the measurement error nature of the problem, however, there is an alternative 'natural' assumption. The variable x_i is said to be a *surrogate* for y_i if the measurement error is *non-differential*, that is if x_i is conditionally independent of (r_i, w_i) given y_i (Carroll, Ruppert and Stefanski, 1995, p.16), or, in a weaker form, if the conditional distribution of x_i given (y_i, w_i, r_i) does not depend upon r_i. Interpreting the conditional distribution of x_i given (y_i, w_i) as the measurement error model, we refer to this assumption as the common measurement error (CME) model assumption, since it implies the measurement error models for respondents and nonrespondents are the same. Note that, under the CME assumption, nonresponse is generally not MAR, since the response r_i may depend on y_i, even conditional on x_i and w_i.

In general, it is not possible to use the data to test the validity of the MAR versus the CME assumptions. One can consider the relative plausibility of both assumptions. One argument in favour of the CME assumption versus the MAR assumption is that it seems more plausible for missingness on y_i to depend directly on the value of y_i than on some surrogate measure of y_i. Nevertheless, like the MAR assumption, the CME assumption is a strong assumption, which may well not hold in reality. For example, it is conceivable that the amount of measurement error for a person who is paid by the hour (and for whom y_i is observed) will be less than for a salaried person for whom y_i is missing. Thus, both assumptions are at best approximations to reality.

The aim of this paper is to develop a method for estimating the distribution of y_i under this CME assumption and to compare it with methods based upon the MAR assumption. We treat the estimation method developed in this paper as a means of assessing the sensitivity of the current method for estimating the distribution of hourly pay to a possible plausible departure from the MAR assumption. It may be argued, as in Manning and Dickens (2002), that MAR-based methods will tend to overestimate numbers of the low paid, if the CME assumption holds, as a result of employees with observed y_i values tending to be lower paid than employees with

missing y_i values. While the direction of the effect may be anticipated, the magnitude of the effect is of some importance for the robustness of MAR-based methods.

Approaches to distribution function estimation, under the CME assumption, include double sampling methods (Luo et al., 1998) and deconvolution methods (Stefanski and Bay, 1996). We do not consider these approaches here, because, for the LFS application, it is unreasonable to assume that the subsample for which y_i is observed may be treated as a simple random subsample or that the measurement error model follows a standard additive form with zero mean and constant variance. Another approach would be to discretise the x_i and y_i variables and to treat this as a misclassification problem, as in Selén (1986). Manning and Dickens (2002) have explored this approach to obtain some estimated upper bounds for low pay proportions, but note that this approach may suffer from small numbers of respondents within the discrete classes.

Our approach will be to consider a data augmentation method, which extends the MAR-based imputation methods considered in Skinner et al. (2002) and Beissel-Durrant and Skinner (2004). Carroll et al. (1995) have proposed similar Gibbs sampling methods to impute the missing values of the variable of interest in the presence of a surrogate variable. However, a replicate variable t_i for all $i \in s$ is required for their approach. Glynn et al. (1993) discuss the use of multiple imputation under nonignorable nonresponse in the presence of follow-up data. However, information on the nonrespondents is required on a randomised subset of the sample. The problem of finding an adequate imputation method is closely related to the problem of parameter estimation of the distribution of the missing values. Kuha (1997) suggests the use of imputation to estimate the parameters of a regression model where some of the variables are subject to measurement error. Again, the presence of validation data is assumed. Some suggestions have been made to use maximum likelihood estimation for the parameters of

interest in the presence of nonignorable nonresponse such as in Greenlees et al. (1982) and Ibrahim and Lipsitz (1996).

The paper is structured as follows. In section 2 the estimation problem for the LFS is described. In section 3, the data augmentation method is developed for inference under the CME assumption. A simulation study is described in section 4, where estimation based on the data augmentation procedure is evaluated, both under the model assumptions made by the method and under forms of misspecification of the model. In section 5 alternative estimates for the LFS are given and compared. Some concluding remarks are made in section 6.

2. Estimating the Distribution of Hourly Pay in the UK

Distributions of hourly pay are important for a wide range of social and economic policy issues in the UK. In particular, to analyse the effects of the National Minimum Wage (NMW), it is crucial to have reliable data about hourly pay, particularly for the bottom end of the pay distribution. We use data from the LFS, a large quarterly survey of households. The quarterly sample is made up of five subsamples, each of about 12,000 addresses. Each quarter one subsample is replaced by a newly selected subsample, designed so that a household remains in the sample for five successive quarterly waves of data collection. Information on earnings is collected in the first and fifth wave, resulting in approximately 16,000 employees each quarter.

We consider two LFS measures of hourly pay. The *derived variable*, obtained by dividing weekly earnings by weekly hours, appears to be subject to appreciable measurement error (Skinner et al. 2002), as in similar surveys in other countries (Rodgers et al., 1993; Moore et al., 2000). The second measure is the *direct variable*. Employees are asked if they are paid by the hour and if yes, they are asked about their (basic) hourly rate. The problem with this variable is not measurement error (which we shall assume to be absent) but missing data. The proportion responding to the direct variable is about 43% overall, with a higher fraction for the lower paid.

Our aim is to estimate the distribution of hourly pay, with particular focus on the bottom end of the distribution, such as the proportion of low paid employees in the population. The cumulative distribution function of hourly pay is defined as

$$F(y) = \frac{1}{N}\sum_{i \in U} I(y_i \leq y), \quad (1)$$

where y_i is the (true) hourly earnings of employee i, U is the population of employees of interest, N is the size of U and $I(.)$ is the indicator function indicating if a condition is true or false and thus $F(y)$ denotes the proportion of employees with hourly earnings not greater than a threshold y. The population may consist of all employees in the UK or of some subpopulation defined by for example age or gender. Only individuals' main jobs will be considered here, ignoring second jobs. The problem is how to estimate $F(y)$, for specified values of y. Following notation in section 1, we let values of the direct variable and derived variable be denoted y_i and x_i respectively for employee i. The response indicator is denoted r_i and a vector of other survey variables denoted w_i. Missingness in x_i and w_i is negligible and we assume here that x_i and w_i are fully observed for all $i \in s$.

To derive a point estimator for $F(y)$, we shall make the simplifying assumption that the population values $(y_i, x_i, w_i, r_i) \sim iid$, i.e. independently and identically distributed (irrespective of whether the employee is in the sample). One possible concern about this assumption is that it takes no account of possible differential unit nonresponse in the LFS. Survey weights have been constructed to address this problem. In principle, weights could be introduced into the point estimation procedure we shall develop, following a pseudo-likelihood procedure (Skinner, 1989) as in Skinner et al. (2002). However, the weights in the LFS do not vary greatly and for simplicity we shall not pursue this approach here. Another possible concern about the *iid* assumption is that it ignores the clustering of individuals in households (the households are not clustered by geography in the LFS design). Ignoring the clustering should not lead to bias in the point

estimators we derive (following the logic of the pseudo-likelihood approach) and is unlikely to lead to more than negligible loss in efficiency (Scott and Holt, 1982). Where it may be important not to ignore the clustering is in variance estimation for the resulting point estimator. Our main interest in this paper is in point estimation and so this issue will only be referred to again briefly in sections 3.4 and 4.2. Note that many low pay estimates are produced separately for men and women, in which case the clustering issue only arises when a household contains more than one person working of a given gender.

Under the *iid* assumption, *if* the y_i were observed completely in s, we could estimate $F(y)$ unbiasedly by

$$\hat{F}(y) = \frac{1}{n}\sum_{i \in s} I(y_i \leq y), \tag{2}$$

where n denotes the number of employees in the sample. The y_i are not, however, fully observed and imputation provides one approach to take account of this problem (Skinner et al., 2002). Imputed values $\hat{y}_i$ are constructed for employees where y_i is missing and $F(y)$ is estimated by

$$\hat{F}_{\cdot}(y) = \frac{1}{n}\sum_{i \in s} I(y_{\cdot i} \leq y), \tag{3}$$

where $y_{\cdot i} = y_i$ if $r_i = 1$ and $y_{\cdot i} = \hat{y}_i$ if $r_i = 0$. In order to obtain an unbiased estimator, we would ideally like to generate the imputed values $\hat{y}_i$ from the conditional distribution of true hourly pay given the derived variable and other covariates. This condition may be expressed as

$$f(y_{\cdot i} \mid x_i, w_i, r_i = 0) = f(y_i \mid x_i, w_i, r_i = 0), \tag{4}$$

where f denotes a generic probability density function. If this condition holds the imputed estimator in (3) would have the same properties as the estimator in (2). The aim is therefore to estimate the distribution $f(y_i \mid x_i, w_i, r_i = 0)$ and to draw imputed values from this estimated

distribution. However, since this distribution cannot be observed directly, further assumptions about the missingness of y_i are required (Little and Rubin, 2002). The MAR assumption referred to in section 1 may be expressed as:

$$y_i \perp r_i \mid x_i, w_i, \tag{5}$$

where $\perp$ denotes independence. Skinner et al. (2002) proposed one imputation method under this assumption, drawing imputed values from the estimated distribution $f(y_i \mid x_i, w_i, r_i = 1)$ leading to an approximately unbiased imputed estimator $\hat{F}.(y)$. The alternative CME assumption referred to in section 1 may be expressed as:

$$x_i \perp r_i \mid y_i, w_i, \tag{6}$$

so that $f(x_i \mid y_i, w_i, r_i = 0) = f(x_i \mid y_i, w_i, r_i = 1)$. The aim of this paper is to develop an imputation method under this CME assumption. This derivation is much less straightforward than under the MAR assumption because the conditioning variables on the right hand side of (6) are subject to missingness.

3. Data Augmentation under the CME Assumption

3.1 Outline of Data Augmentation Approach

Data augmentation is a Markov chain Monte Carlo method, which enables imputation for complex missing data problems by iteratively solving more tractable complete data problems (Schafer, 1997 and Gelman et al., 1998). The method is most naturally viewed from a Bayesian perspective, although the resulting imputed values can be used for frequentist purposes as in our application. In the context of missing data, the data augmentation algorithm consists of a series of *imputation steps* (*I*-steps), which impute the missing values given all the observed data and a current set of parameters, and *posterior steps* (*P*-steps), in which the parameters of the model are

drawn from their posterior distribution given the complete data formed from the *I*-step. On convergence, the algorithm should provide imputed values from the conditional distribution of the missing values given the observed data, as in (4), where the distribution is integrated over any unknown parameters in the model with respect to the posterior distribution of these parameters given the data. To apply data augmentation to our problem we must first formulate our model more fully. We shall then specify the imputation step and posterior step in sections 3.2 and 3.3 respectively.

We introduce the following notation. The vectors of length n, containing the sample values are denoted Y, X and R, for example $Y = (y_1,...,y_n)'$. Similarly, W denotes a matrix with values of the covariates. We suppose without loss of generality that for the direct variable only the first n_r elements are observed in sample s and the following $n - n_r$ elements are missing. We write $Y = (Y'_{obs}, Y'_{mis})'$, where $Y_{obs} = (y_1,...,y_{n_r})'$ is the observed part of Y and $Y_{mis} = (y_{n_r+1},...,y_n)'$ is the missing part.

For our application, we consider a model for Y, X, R conditional on W which we express as $f(Y, X \mid W, \zeta) f(R \mid Y, X, W, \psi)$, where ζ and ψ are the parameters of the complete data and the missing data mechanism respectively. We shall also require a prior density $f(\zeta, \psi)$ for ζ and ψ. The predictive distribution of the direct variable required for the *I*-step is $f(Y \mid X, R, W, \zeta, \psi)$ and the complete-data posterior required for the *P*-step is $f(\zeta, \psi \mid Y, X, R, W)$. It is convenient to express the parameter ζ as $(\zeta'_1, \zeta'_2)'$, where ζ_1 is the vector of parameters of $f(Y \mid X, W, \zeta_1)$ and ζ_2 is the vector of parameters of $f(X \mid W, \zeta_2)$. Using the CME assumption, we have the factorisation

$$f(Y, X, R \mid W, \zeta, \psi) = f(Y \mid X, W, \zeta_1) f(X \mid W, \zeta_2) f(R \mid Y, W, \psi), \qquad (7)$$

which is convenient for the implementation of the *I*- and the *P*-steps. This factorisation into three models has a simple interpretation. The first model represents the predictive distribution

of the true hourly pay, the second the predictive distribution of the variable measured with error and the third factor represents a model for the nonresponse under the CME assumption.

3.2 The Imputation Step

The imputation step requires drawing imputed values for missing values of y_i from the predictive distribution $f(y_i \mid x_i, w_i, r_i = 0, \zeta, \psi)$. Using the property under CME and (7), we have

$$f(y_i \mid x_i, w_i, r_i = 0, \zeta, \psi) = \frac{f(y_i, x_i, r_i = 0 \mid w_i, \zeta, \psi)}{f(x_i, r_i = 0 \mid w_i, \zeta, \psi)} = f(y_i \mid x_i, w_i, \zeta_1) \frac{f(r_i = 0 \mid y_i, w_i, \psi)}{f(r_i = 0 \mid x_i, w_i, \zeta, \psi)}$$

and therefore

$$f(y_i \mid x_i, w_i, r_i = 0, \zeta, \psi) \propto f(y_i \mid x_i, w_i, \zeta_1) f(r_i = 0 \mid y_i, w_i, \psi). \tag{8}$$

The *I*-step may thus be implemented as follows. Given current values of the parameters $\zeta_1^{(d)}$ and $\psi^{(d)}$, where d denotes the iteration of the data augmentation procedure, $d = 0, \ldots, D$, a possible imputed value for nonrespondent i, denoted $\hat{y}_i^{(d+1)*}$, is drawn $\hat{y}_i^{(d+1)*} \sim f(y_i \mid x_i, w_i, \zeta_1^{(d)})$. Rejection sampling (Tanner, 1996 and Gelman et al., 1998) is then performed based on the nonresponse model, accepting $\hat{y}_i^{(d+1)*}$ for imputation with probability $f(r_i = 0 \mid \hat{y}_i^{(d+1)*}, w_i, \psi^{(d)}) = \rho_i^{(d+1)*}$, where $\rho_i^{(d+1)*}$ denotes the probability of nonresponse. If accepted, we set $\hat{y}_i^{(d+1)*} = \hat{y}_i^{(d+1)}$, where $\hat{y}_i^{(d+1)}$ is the imputed value for nonrespondent i at iteration $d+1$. If rejected, another value $\hat{y}_i^{(d+1)*}$ is drawn and so on. The *I*-step in (8) has therefore a simple interpretation and is easy to implement. The model $f(y_i \mid x_i, w_i, \zeta_1)$ is henceforth referred to as the *imputation model* and $f(r_i = 0 \mid y_i, w_i, \psi)$ as the *nonresponse model*. An advantage of the factorisation in (7) is that a model for X does not need to be fitted, and therefore no assumptions need to be made about this distribution.

To draw values $\hat{y}_i^{(d+1)*}$ from $f(y_i \mid x_i, w_i, \zeta_1^{(d)})$ in practice, we initially use a standard parametric regression model. Using the logarithmic transformation for y_i it is assumed that

$$\ln(y_i) \mid x_i, w_i, \zeta_1 \sim N(\eta_i \beta; \sigma^2_{Y|X,W}), \tag{9}$$

where η_i is a vector of covariates, functions of x_i and w_i, β is a vector of coefficients and $\sigma^2_{Y|X,W}$ denotes the conditional variance of $\ln(y_i)$ given x_i and w_i. The vector of parameters is $\zeta_1 = (\beta', \sigma^2_{Y|X,W})'$. Regression imputation can then be performed adding a normal error to the predicted values from the model (David et al. 1986). Similar assumptions as in (9) have been made by Raghunathan et al. (2001) and by Heitjan and Rubin (1990) for heaped, rounded and truncated data. Greenlees et al. (1982) use these assumptions for estimating regression models for income data in the presence of nonignorable nonresponse and find that the assumptions approximately hold for an earnings variable obtained from a validation study. Since the assumptions in (9) refer to respondents and nonrespondents, the validity of these distributional assumptions are strictly speaking untestable for our example. Applying the model in (9) to LFS data based on respondents only, i.e. where $r_i = 1$, we found that the residuals are approximately normally distributed. However, there was an indication that the residual variance may increase with increasing predicted values such that the assumption of homoscedasticity may not be adequate.

To relax the distributional assumptions made in (9), in particular to address the problem of a possible departure from the assumption of constant variance, the use of hot deck imputation instead of parametric regression imputation is considered for the *I*-step. Predictive mean matching imputation (Little, 1988 and Heitjan and Little, 1991), which is based on the predictions for $\ln(y_i)$, is proposed. This form of donor imputation has the advantage that actually observed values are imputed that may preserve the shape of the earnings distribution, for example that preserve truncation, heaping and rounding effects. Two forms of predictive mean matching imputation are implemented, hot deck imputation within classes and nearest neighbour imputation, where the classes and the nearest neighbours are defined based on the predictions of the regression model for $\ln(y_i)$. The imputation classes are defined as equally

spaced intervals of the range of the predicted values, such as £1.50 classes. In total 9 classes were used. Under hot deck imputation within classes Q donor values, denoted $\hat{y}_{i1}^{*},...,\hat{y}_{iQ}^{*}$, are selected with simple random sampling without replacement for nonrespondent i from the same class. Under nearest neighbour imputation the $Q/2$ responding nearest neighbours above and below the predicted value for nonrespondent i are used to obtain the Q possible values for imputation, where the value for Q is an even number. However, under hot deck imputation within classes and nearest neighbour imputation the number of values that can be chosen for imputation is restricted due to the definition of the classes and the nearest neighbours. The acceptance-rejection procedure based on the probability $\rho_i = 1 - f(r_i = 1 \mid y_i, w_i, \psi)$ is therefore modified using a weighted bootstrap method as described in Carroll et al. (1995) and Tanner (1996), since classical rejection sampling requires being able to generate a large number of potential imputed values, which is only possible under parametric random regression imputation. Under the weighted bootstrap method the value for imputation, $\hat{y}_i^{(d+1)}$, for iteration $d+1$, is sampled out of the Q possible values $\hat{y}_{i1}^{(d+1)*},...,\hat{y}_{iQ}^{(d+1)*}$ with probabilities

$$\breve{\rho}_{iq}^{(d+1)^{*}} = f(r_i = 0 \mid y_{iq}^{(d+1)*}, w_i, \psi^{(d)}) / \sum_{q=1}^{Q} f(r_i = 0 \mid y_{iq}^{(d+1)*}, w_i, \psi^{(d)}), \tag{10}$$

for all $q = 1,...,Q$. Note that under both, rejection sampling and the weighted bootstrap method, in each *I*-step only one value $\hat{y}_j^{(d+1)}$ is imputed for each nonrespondent. The proposed *I*-step extends the MAR-based imputation procedure described in Skinner et al. (2002) where values are drawn from the predictive distribution $f(y_i \mid x_i, w_i, r_i = 1)$ without the addition of rejection sampling.

3.3 The Posterior Step

The *P*-step requires drawing values of the parameters from the complete data posterior distribution. Under the factorisation (7) this distribution can be expressed as

$$f(\zeta_1,\zeta_2,\psi \mid Y,X,R,W) \propto f(Y \mid X,W,\zeta_1)f(\zeta_1)\, f(X \mid W,\zeta_2)f(\zeta_2)\, f(R \mid Y,W,\psi)f(\psi), \tag{11}$$

assuming that the prior distribution $f(\zeta_1,\zeta_2,\psi)$ factors into $f(\zeta_1)f(\zeta_2)f(\psi)$, i.e. that the parameters ζ_1, ζ_2 and ψ are *a priori* independent. Since the likelihood function also factorises with respect to these parameters, the posterior distribution of ζ_1, ζ_2 and ψ factorises into three independent posteriors. A posterior for ζ_2 does not need to be specified since the *I*-step does not require a model for X.

To proceed, we need to compute (11) specifying the likelihood function and priors. An assumption about $f(Y \mid X,W,\zeta_1)$ was already made in (9). For the response model the following notation is introduced. Let $f(r_i = 1 \mid y_i, w_i, \psi) = G(\tau_i\psi) = p_i$, where τ_i is a row-vector including functions of y_i and w_i, p_i denotes the probability of response and G the logistic regression model, $G(\tau_i\psi) = \dfrac{\exp(\tau_i\psi)}{1+\exp(\tau_i\psi)}$. It follows that the complete-data likelihood, assuming independence across units given the parameters, is:

$$\begin{aligned} f(Y,X,R \mid W,\zeta,\psi) \propto & \prod_{i=1}^{n}(\sigma^2_{Y|X,W})^{-1/2}\exp\left\{\frac{-1}{2\sigma^2_{Y|X,W}}\,(\ln(y_i) - \eta_i\beta)^2\right\} \\ & * \prod_{i=1}^{n} f(x_i \mid w_i,\zeta_2) * \prod_{i=1}^{n} G^{r_i}(\tau_i\psi)\{1-G(\tau_i\psi)\}^{(1-r_i)}, \end{aligned} \tag{12}$$

where $\eta_i\beta = E(\ln(y_i) \mid x_i, w_i)$.

We now turn to the specification of the prior distributions. Given the large dataset and the lack of any clear prior information about the parameters, we seek computationally convenient noninformative priors (Box and Tiao, 1992 and Gelman et al., 1998). A noninformative prior for ζ following calculations in Schafer (1997) and Box and Tiao (1992) is

$$f(\zeta) \propto (\sigma^2_{X|W})^{-1/2}(\sigma^2_{Y|X,W})^{-3/2}, \tag{13}$$

where $\sigma^2_{X|W}$ denotes the conditional variance of X given W, so that the prior for ζ factors into independent priors for ζ_1 and ζ_2. The prior in (13) is derived using a noninformative prior for the parameter of the joint distribution of X and Y given W, representing a limiting form of the normal inverted Wishart density, and applying the one-to-one transformation between the parameter of this joint distribution and ζ. The resulting posterior distribution for $\zeta_1 = (\beta', \sigma^2_{Y|X,W})'$, discarding proportionality constants, is

$$f(\zeta_1 \mid Y, X, W) \propto (\sigma^2_{Y|X,W})^{-3/2} \prod_{i=1}^{n} (\sigma^2_{Y|X,W})^{-1/2} \exp\left\{ \frac{-1}{2\sigma^2_{Y|X,W}} (\ln(y_i) - \eta_i \beta)^2 \right\}$$

$$= (\sigma^2_{Y|X,W})^{-(\frac{n+3}{2})} \exp\left\{ \frac{-1}{2\sigma^2_{Y|X,W}} \sum_{i=1}^{n} (\ln(y_i) - \eta_i \beta)^2 \right\}. \qquad (14)$$

In the special case that the data has a monotone missing-data pattern (Little and Rubin, 2002), as in our example, and since the parameters are independent, the posterior $f(\zeta_1 \mid Y, X, W)$, following derivations in Box and Tiao (1992), can be expressed as the product of a multivariate normal distribution, $N(\hat{\beta}, \sigma^2_{Y|X,W}(\eta'\eta)^{-1})$, and a scaled inverted chisquare distribution, $\hat{\varepsilon}'_Y \hat{\varepsilon}_Y \chi^{-2}_{n-1}$, with *n*-1 degrees of freedom and scaling factor $\hat{\varepsilon}'_Y \hat{\varepsilon}_Y$, where β and η_i are defined in (9) and η defines the corresponding matrix, $\hat{\beta}$ is the maximum likelihood estimate, $\hat{\beta} = (\eta'\eta)^{-1}\eta \ln(Y)$, and $\hat{\varepsilon}_Y = \ln(Y) - \eta\hat{\beta}$, both based on augmented data $\ln(Y)$. The required parameters can therefore be drawn from the posterior distribution as follows

$$\sigma^2_{Y|X,W} \mid Y, X, W \sim \hat{\varepsilon}'_Y \hat{\varepsilon}_Y \chi^{-2}_{n-1} \quad \text{and} \quad \beta \mid \sigma^2_{Y|X,W}, Y, X, W \sim N(\hat{\beta}, \sigma^2_{Y|X,W}(\eta'\eta)^{-1}). \qquad (15)$$

We now turn to the problem of drawing parameters for the response model $f(R \mid Y, W, \psi)$ in the posterior step based on complete data. Several approaches for specifying priors for binomial regression problems have been proposed (Bedrick et al., 1996). It is common to assume a normal or noninformative prior $f(\psi)$, which is convenient in large sample situations where the

posterior of ψ is approximately normal (Zellner and Rossi, 1984). Here, the prior is specified as $f(\psi) \propto c$, where c is a constant, such that for the complete-data posterior of ψ we have

$$f(\psi \mid Y,R,W) \propto f(R \mid Y,W,\psi)c \propto f(R \mid Y,W,\psi) = \exp(\log(f(R \mid Y,W,\psi))). \tag{16}$$

Following derivations in Zellner and Rossi (1984) and expanding $L(\psi) \equiv \log(f(R \mid Y,W,\psi))$ in a Taylor series about the modal value of (16), i.e. the maximum likelihood estimate $\hat{\psi}$, and using the first order approximation, we obtain

$$f(\psi \mid Y,R,W) \propto \exp\{-\frac{1}{2}(\psi - \hat{\psi})'T(\psi - \hat{\psi})\}, \tag{17}$$

such that ψ follows approximately a multivariate normal distribution with mean $\hat{\psi}$ and variance-covariance matrix T^{-1},

$$\psi \sim N(\hat{\psi},T^{-1}). \tag{18}$$

The matrix T is defined as $T = -\left[\frac{\partial^2 L(\psi)}{\partial\psi\partial\psi'}\right]_{\psi=\hat{\psi}} = \tau'V\tau$, where τ is a matrix including functions of Y and W and V is a diagonal matrix with element

$$v_i = \left[\frac{r_i}{G_i^2} + \frac{1-r_i}{(1-G_i)^2}\right]g_i^2 - \frac{(r_i - G_i)g_i'}{G_i(1-G_i)}, \tag{19}$$

where $G_i = G(\tau_i\hat{\psi})$, $g_i = \left[\frac{dG(\mathrm{z}_i)}{d\mathrm{z}_i}\right]_{\mathrm{z}_i=\tau_i\hat{\psi}} = g(\tau_i\hat{\psi})$ and $g_i' = \left[\frac{dg(\mathrm{z}_i)}{d\mathrm{z}_i}\right]_{\mathrm{z}_i=\tau_i\hat{\psi}}$. We now have specified the required imputation and posterior step for data augmentation under the CME assumption.

3.4 Inference Under Data Augmentation

Suppose that the data augmentation algorithm has run long enough to achieve approximate stationarity and to be independent of the initial starting values $\zeta_1^{(0)}$ and $\psi^{(0)}$, i.e. d is large enough such that the vectors of parameters $\zeta_1^{(d)}$ and $\psi^{(d)}$ are essentially draws from the observed-data posterior. Then the imputed values of y_i will follow the distribution in (4) and the estimator in (3) will be approximately unbiased, under the model assumptions. In order to

improve the efficiency of this estimator, $M > 1$ values $\hat{y}_i^{(m)}$, $m = 1,...,M$, may be determined for each nonrespondent from repeated *I*-steps. The resulting point estimators from each of the M completed datasets, denoted $\hat{F}_{\cdot}^{(m)}(y)$ for $m = 1,...,M$, may then be combined (Rubin, 1987) to give the point estimator:

$$\hat{F}_{\cdot}(y) = \frac{1}{M}\sum_{m=1}^{M}\hat{F}_{\cdot}^{(m)}(y). \tag{20}$$

The method of multiple imputation, moreover, suggests a method of variance estimation in the context of data augmentation (Little and Rubin, 2002). For the purpose of variance estimation, the M sets of multiple imputations should not be obtained from successive sets of imputed values Y_{mis} since they are correlated. Instead, the Markov chain may be subsampled after an initial burn-in period using every k-th iterate to achieve approximate independence of repeated imputations (see section 4.1 for choice of k). An estimator of the variance of $\hat{F}_{\cdot}(y)$ is then given by (Rubin, 1987):

$$\hat{\mathrm{var}}_{MI}(\hat{F}_{\cdot}(y)) = \bar{A}_{\cdot} + (1 + 1/M)\hat{B}_{\cdot}\ , \tag{21}$$

where $\bar{A}_{\cdot} = \frac{1}{M}\sum_{m=1}^{M}\hat{A}_{\cdot}^{(m)}$ is the within imputation variance, and $\hat{A}_{\cdot}^{(m)}$ is the standard variance estimator valid for complete data, applied to Y_{obs} and the imputed values $Y_{mis}^{(m)}$ for the m-th imputation, and $\hat{B}_{\cdot} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{F}_{\cdot}^{(m)}(y) - \hat{F}_{\cdot}(y))^2$ is the between imputation variance. This variance estimator fails, however, to take account of the clustering in the LFS.

4. Simulation Study

4.1 Design of the Simulation Study

The aim of this study is to evaluate the performance of the point estimator (20) empirically under ideal conditions and under misspecification of the imputation and the nonresponse

model. The main emphasis is on the bias of the point estimator. Data augmentation under CME is compared to MAR-based imputation methods. Independent repeated samples $s^{(h)}$, $h = 1,...,H,$ are generated with values $y_i, x_i, w_i, r_i, \quad i \in s^{(h)}$. To reflect the features of the LFS, values of w_i are generated from data for approximately 16,000 employees in the March-May 2000 LFS quarter using simple random sampling with replacement, i.e. adopting a bootstrap approach. The bootstrap approach provides flexibility in the choice of sample size, while treating the underlying population as infinite, in line with the small sampling fraction of the LFS. Variables that are likely to be predictors of hourly earnings, measurement error in the derived variable or nonresponse are included in w_i. The values for y_i, x_i and r_i are generated from a model for different reasons. The values y_i are modelled since this variable is subject to missing data in the original LFS – $\ln(y_i)$ is generated from a fitted linear regression on $\ln(x_i)$ and six other covariates including squared terms and with an added normal error. The values r_i are modelled so that the nonresponse (missing data) mechanism takes various known forms, including CME, MAR and a nonignorable form that is neither CME nor MAR. Logistic regression models are used relating r_i to $\ln(y_i)$, $\ln(x_i)$ and other covariates that are likely to be predictors of nonresponse. The values for x_i are simulated from a model to avoid duplications of values (x_i, w_i) in $s^{(h)}$. Duplications of units were regarded as unrealistic since the proposed imputation methods are based on the predicted values for y_i given (x_i, w_i) and duplications would have led to identical predicted values for some units $i \in s^{(h)}$. The model generating $\ln(x_i)$ has six covariates including squared terms and a normal error. Some of the covariates are also included in the model which generates $\ln(y_i)$. Note that the covariates involved in the three models for $\ln(y_i)$, $\ln(x_i)$ and r_i are not necessarily the same since predictions of different types of variables are required. All models were fitted to respondents in the original LFS sample to obtain estimates of the required parameters.

Ideally the sample $s^{(h)}$ should be of the same size as the original LFS sample but, because of the computer intensive nature of the data augmentation approach, this is reduced to $n = 1000$ with the number of simulation replications being set at $H = 100$. The specifications for the data augmentation procedure are as follows. A single Markov Chain was generated and convergence of the algorithm was tested using time series analysis based on components of ζ and ψ. To determine the subsampling constant k (see section 3.4), autocorrelation functions were investigated as described in Schafer (1997). An initial burn-in period of 200 iterations and a subsampling constant of $k = 100$ resulting in an overall length of $D = 1100$ were found adequate. The initial starting values $\zeta_1^{(0)}$ and $\psi^{(0)}$ are the maximum likelihood estimates based on respondent data in $s^{(h)}$. Alternative specifications of the burn-in period, the parameters k and D, as well as other starting values $\zeta_1^{(0)}$ and $\psi^{(0)}$ were used, however, leading to very similar results. The total number of imputations for all imputation methods used here is $M = 10$.

The performances of the following three point estimators are investigated: $\hat{F}_1.$ is the estimated proportion of employees paid below the NMW (=£3.00 per hour aged 18-21, £3.60 per hour aged 22+ in March 2000), $\hat{F}_2.$ is the estimated proportion paid at the NMW (5p above and below the threshold) and $\hat{F}_3.$ is the estimated proportion paid between the NMW and £5 per hour. The 'true' values determined via simulation are $F_1 = 0.95\%$, $F_2 = 10.41\%$ and $F_3 = 27.41\%$. The simulation estimates of bias and standard errors are defined as $\hat{bias}(\hat{F}.) = \bar{F}. - F$ and $\hat{se}(\hat{F}.) = [(H-1)^{-1}\sum_{h=1}^{H}(\hat{F}.^{(h)} - \bar{F}.)^2]^{1/2}$, where $\bar{F}. = H^{-1}\sum_{h=1}^{H}\hat{F}.^{(h)}$. Data augmentation as set out in section 3 under the CME assumption is referred to as DA-CME. The abbreviations Reg Imp, NN and IC refer to regression imputation, nearest neighbour and hot deck imputation based on imputation classes respectively. In addition, standard data augmentation based on the MAR assumption without the addition of rejection sampling, as for example described in Schafer (1997), is implemented and is referred to as DA-MAR. For

comparison, MAR-based imputation without draws of parameters from the corresponding posterior distributions and without rejection sampling as described in Beissel-Durrant and Skinner (2004) is analysed.

4.2 Results of the Simulation Study

We first analyse the performance of DA-CME under a correct CME nonresponse mechanism and correct covariates so that the imputation model coincides with the model generating $\ln(y_i)$. The results are presented in Table 1. Different imputation methods are compared. Regression imputation performs as expected very well with no significant bias. Nearest neighbour imputation also performs well with all biases below 2%. The significant bias for $\hat{F}_3.$ for nearest neighbour imputation might be caused by the use of the weighted bootstrap, which is an approximation to the rejection sampling procedure. To improve the results for NN the parameter Q may be increased from 10 to 20 which was found to lead to an improvement in the result for $\hat{F}_2.$ and $\hat{F}_3..$ Since the results are very similar, however, only $Q = 10$ was used in the following. For hot deck imputation within classes all estimators are not significantly biased. An increased value for Q leads as expected to a reduction in the bias. However, a higher relative bias was obtained for $\hat{F}_1.$ which is thought to be related to the definition of the imputation classes and indicates a potential disadvantage of the IC method, since the classes are defined arbitrarily and the performance of the point estimators may depend on the definition of the classes. Hot deck imputation within classes is therefore not analysed any further here. We conclude that there is no evidence of important bias for the DA-CME method if the imputation model and the nonresponse model are correctly specified.

[Table 1 about here]

Table 2 provides estimates of corresponding simulation standard errors for the nearest neighbour imputation method using DA-CME. The results are compared to standard errors of

the three point estimators derived under data augmentation based on the MAR assumption (DA-MAR) and MAR-based NN imputation. As expected, we have a slightly higher estimate for DA-CME than for the MAR-based methods, caused by the additional draws from the posterior distributions of the parameters and the use of rejection sampling in the imputation step. However, the increase is around or less than 10%. To investigate the performance of the multiple imputation variance estimation formula, simulation estimates of biases and coverage rates are given in Table 3. Only $\hat{\text{var}}_{MI}(\hat{F}_3.)$ for NN imputation was found to be significantly biased. The estimated coverage rates are close to 95% for the 95% confidence intervals. Thus, the multiple imputation variance formula performs reasonably well for DA-CME. Note that these results are only based on $H = 100$ iterations. For a more detailed analysis $H = 1000$ or higher is recommended. Note that the multiple imputation formula does not currently allow for clustering of individuals within households.

[Table 2 and 3 about here]

Of particular interest is the performance of the DA-CME method under misspecification of the imputation model and the nonresponse model. Table 4 summarises the performance of the method under misspecification of the imputation model. The misspecifications 1-3 indicate an increasingly more complex model generating y_i in comparison to the assumed imputation model. Misspecification 3 in addition allows for a model for x_i that differs significantly from the model that generates y_i. The case where the values x_i are not generated but original values from the LFS sample are used via the bootstrap approach is also investigated. As expected, with an increasing degree of misspecification the amount of bias for all three point estimators increases with some biases being significant. This amount is greatest for regression imputation for $\hat{F}_1$. which may reflect a greater sensitivity of the regression imputation procedure to misspecification of the imputation model. In comparison, nearest neighbour hot deck imputation seems to perform reasonably well under misspecification of the imputation model with almost all

estimated relative biases below 3%. It should be noted that DA-CME may not fully rely on the specifications of the imputation model since this model is used to generate possible values $\hat{y}_i^*$ which are then accepted or rejected for imputation by the nonresponse model. This can be seen for example in the generally good performance of $\hat{F}_2.$ and $\hat{F}_3.$ under misspecification of the imputation model. Here, nearest neighbour imputation seems to be less dependent on assumptions about the model and may therefore be preferable to regression imputation.

[Table 4 about here]

Since the DA-CME method also requires the specification of a nonresponse model it is of interest to analyse its performance under misspecification of the nonresponse mechanism. Table 5 shows the results of DA-CME when in fact a nonignorable nonresponse mechanism holds that is a.) an extended version of CME which includes 6 more covariates in the model generating nonresponse (CME+6), and b.) dependent on x_i in addition to y_i and w_i and therefore dependent on all variables in the simulated dataset, which is referred to as the full model. The case of a MAR nonresponse mechanism is also considered. As expected, for almost all estimators we observe an increase in the bias. Regression imputation seems to be more sensitive to misspecification of the nonresponse model than nearest neighbour imputation leading to some significant biases of around 6-8% for $\hat{F}_2.$ and $\hat{F}_3.$ under the full model and MAR nonresponse. This is thought to be related to the use of rejection sampling where the specification of the nonresponse model is of direct relevance. The weighted bootstrap as used in the hot deck imputation method seems to be less dependent on parametric assumptions about the nonresponse model. For nearest neighbour imputation most of the biases are below 3% for all three nonresponse mechanisms. We conclude that DA-CME using nearest neighbour imputation seems reasonably robust to misspecification of the nonresponse mechanism and performs well even under MAR and a full nonignorable nonresponse mechanism. In comparison, Table 6 shows the performance of the MAR-based imputation methods under the

CME nonresponse mechanism. The results seem to indicate a slightly higher sensitivity to misspecification of the nonresponse model for the MAR-based methods than for DA-CME. However, the overestimation seems to be less than 7% which does not make the resulting estimates unusable. Here, regression imputation and nearest neighbour imputation seem to perform very similarly if the imputation model is correctly specified.

[Table 5 and 6 about here]

5. Application to the UK Labour Force Survey

The different imputation methods are also applied to the Labour Force Survey. Table 7 shows estimates $\hat{F}_{1.}$, $\hat{F}_{2.}$ and $\hat{F}_{3.}$ for the March-May 2000 quarter. We can see that as expected the DA-CME methods lead to a reduction of the estimates in comparison to the MAR-based methods. For $\hat{F}_{1.}$, the estimated proportion of employees earning below the NMW, we have 0.45% for DA-CME whereas for the MAR-based methods we obtain a value of 0.50%, which indicates an overestimation of about 10% by the MAR-based methods. This coincides approximately with findings from the simulation study which indicated an overestimation of less than 7%. We find that DA-MAR and the MAR-based methods lead as expected to very similar estimates. The data augmentation procedures were also analysed using different starting values for the unknown parameters as well as different run times D and subsampling constants k leading to similar results as in Table 7.

For all three methods DA-CME, DA-MAR and MAR-based imputation nearest neighbour and hot deck imputation within classes seem to produce very similar estimates. Overall, random regression imputation did not seem to perform very well in the actual application to LFS data. The estimate $\hat{F}_{1.}$ under regression imputation for example is higher than for the two hot deck imputation methods (1.23% in comparison to 0.50% in the case of MAR-based imputation). This is thought to be related to a greater dependency of the regression method on distributional

assumptions, in particular the assumption of constant variance in the imputation model. We therefore applied regression imputation where the variance of the added on residuals is defined within classes of the predicted values, i.e. allowing for a non-constant variance. We observed that the estimates $\hat{F}_1$. for example were reduced, which indicates a stronger dependency on distributional assumptions under regression imputation. It was found that the hot deck methods seem to preserve certain features of the hourly pay distribution better, in particular truncation and rounding effects. Overall, nearest neighbour hot deck imputation seems to be preferable.

[Table 7 about here]

6. Conclusions

If ignored, measurement error may lead to serious bias in the estimation of hourly pay distributions, particularly at the lower end. In this paper, we have considered alternative estimation methods, which correct for measurement error using a subsample of accurately measured values of hourly pay. Existing methods assume that these accurately measured values are missing at random (MAR). We have developed a new estimation method under an alternative nonignorable missingness assumption, that a common measurement error (CME) process applies to respondents and nonrespondents. The method adapts the Bayesian method of data augmentation using hot deck imputation to allow for the 'spiky' nature of the hourly pay data, which does not follow a simple homoscedastic parametric regression model. Our simulation study shows that the method produces approximately unbiased estimates under correct specification of models and that the method is reasonably robust against misspecification of the imputation model and of the nonresponse model (in the case of nearest neighbour hot deck imputation). In particular, the method showed a good performance under a general nonignorable nonresponse mechanism.

Existing MAR-based methods are found to overestimate proportions of low paid employees by no more than 10%, in relative terms, in the simulation study. Using LFS data from March-May 2000, the existing MAR-based method, similar to that used by the Office for National Statistics, estimates the proportion of employees earning below the National Minimum Wage (£3.00 per hour aged 18-21 and £3.60 per hour aged 22+) as 0.50%, compared with an estimate of 0.45% for the new CME-based method (using nearest neighbour imputation). This suggests a similar degree of overestimation by the MAR-based method when applied to LFS data, although further work is needed on standard error estimation. The new method displays somewhat higher standard errors than the MAR-based methods (e.g. 10% higher for some low pay estimates). Some variance estimation methods have been developed for the new method based on multiple imputation formulae, but these are tentative and need extending for the LFS data to allow, in particular, for clustering of individuals within households.

References

Bedrick, E.J., Christensen, R. and Johnson, W. (1996) A New Perspective on Priors for Generalised Linear Models, *Journal of the American Statistical Association*, **91**, 1450-1460.

Beissel-Durrant, G. and Skinner, C.J. (2004) Estimation of the Distribution of Hourly Pay from Household Survey Data: The Use of Missing Data Methods to Handle Measurement Error, *Methodology Working Paper Series*, M04/08, available from: http://www.s3ri.soton.ac.uk/publications/methodology.php.

Box, G.E.P. and Tiao, G.G. (1992) *Bayesian Inference in Statistical Analysis*, Reading, 1992.

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995) *Measurement Error in Nonlinear Models*, London, Chapman and Hall.

David, M., Little, R.J.A., Samuhel, M.E. and Triest, R.K. (1986) Alternative Methods for CPS Income Imputation, *Journal of the American Statistical Association*, **81**, 29-41.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1998) *Bayesian Data Analysis*, London.

Glynn, R.J., Laird, N.M. and Rubin, D.B. (1993) Multiple Imputation in Mixture Models for Nonignorable Nonresponse with Follow-ups, *Journal of the American Statistical Association*, **88**, 423, 984-993.

Greenlees, J.S., Reece, W.S., Zieschang, K.D. (1982) Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed, *Journal of the American Statistical Association*, **77**, 251-261.

Heitjan, D.F. and Little, R. (1991) Multiple Imputation for the Fatal Accident Reporting System, *Journal of the Royal Statistical Society*, *Applied Statistics*, **40**, 1, 13-29.

Heitjan, D.F. and Rubin, D.B. (1990) Inference from Coarse Data via Multiple Imputation with Application to Age Heaping, *Journal of the American Statistical Association*, **85**, 410, 304-314.

Ibrahim and Lipsitz (1996) Parameter Estimation from Incomplete Data in Binomial Regression when the Missing Mechanism is Nonignorable, *Biometrics*, **52**, 1071-1078.

Kuha, J. (1997) Estimation by Data Augmentation in Regression Models with Continuous and Discrete Covariates Measured with Error, *Statistics in Medicine*, **16**, 189-201.

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

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data*, New York, Wiley.

Luo, M., Stokes, L. and Sager, T. (1998) Estimation of the CDF of a Finite Population in the Presence of a Calibration Sample, *Environmental and Ecological Statistics*, **5**, 277-289.

Manning, A. and Dickens, R. (2002) *The Impact of the National Minimum Wage on the Wage Distribution, Poverty and the Gender Pay Gap,* Working Paper prepared for the Low Pay Commission, 1-98; available from:
http://www.lowpay.gov.uk/lowpay/research/pdf/amrd.pdf

Moore, J.C., Stinson, L.L. and Welniak, J.E. (2000) Income Measurement Error in Surveys, A Review, *Journal of Official Statistics*, **16**, 331-361.

Raghunathan, T.E., Lepkowski, J.M., Hoewyk, J.V. and Solenberger, P. (2001) A Multivariate Technique for Multiply Imputing Missing Values using a Sequence of Regression Models, *Survey Methodology*, **27**, 1, 85-95.

Rodgers, W.L., Brown, C. and Duncan, G.J. (1993) Errors in Survey Reports of Earnings, Hours Worked and Hourly Wages, *Journal of the American Statistical Association*, **88**, 1208-1218.

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

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

Scott, J. and Holt, D. (1982) The Effect of Two-Stage Sampling on Ordinary Least Squares Methods, *Journal of the American Statistical Association*, **77**, 848-854.

Selén, J. (1986) Adjusting for Errors in Classification and Measurement in the Analysis of Partly and Purely Categorical Data, *Journal of the American Statistical Association*, **81**, 75-80.

Skinner, C., Stuttard, N., Beissel-Durrant, G. and Jenkins, J. (2002) The Measurement of Low Pay in the U.K. Labour Force Survey, *Oxford Bulletin of Economics and Statistics*, **64**, 653-676.

Skinner, C.J. (1989) Domain Means, Regression and Multivariate Analysis, in C.J. Skinner, D. Holt and T.M.F. Smith (eds.), *Analysis of Complex Surveys*, Chichester, Wiley.

Stefanski, L.A. and Bay, J.M. (1996) Simulation Extrapolation Deconvolution of Finite Population Cumulative Distribution Function Estimators, *Biometrika*, **83**, 407-417.

Tanner, M.A. (1996) *Tools for Statistical Inference, Methods for the Exploration of Posterior Distributions and Likelihood Functions,* Springer, New York.

Zellner, A. and Rossi, P.E. (1984) Bayesian Analysis of Dichotomous Quantal Response Models, *Journal of Econometrics*, 365-393.

Table 1: Simulation Estimates of Biases of Estimators $\hat{F}_1.$, $\hat{F}_2.$ and $\hat{F}_3.$ for Different Imputation Methods using Data Augmentation DA-CME, Assuming CME and Correct Covariates.

DA-CME	Bias of $\hat{F}_1.$	Rel. Bias of $\hat{F}_1.$	Bias of $\hat{F}_2.$	Rel. Bias of $\hat{F}_2.$	Bias of $\hat{F}_3.$	Rel. Bias of $\hat{F}_3.$
Reg Imp	$0.10*10^{-3}$ ($0.29*10^{-3}$)	1.05 %	$0.46*10^{-3}$ ($0.94*10^{-3}$)	0.44 %	$0.60*10^{-3}$ ($1.34*10^{-3}$)	0.21 %
NN, Q=10	$0.06*10^{-3}$ ($0.34*10^{-3}$)	-0.65 %	$1.19*10^{-3}$ ($1.13*10^{-3}$)	1.14 %	$5.11*10^{-3}$ ($1.49*10^{-3}$)	1.86 %
NN, Q=20	$-0.15*10^{-3}$ ($0.34*10^{-3}$)	-1.62 %	$0.97*10^{-3}$ ($1.11*10^{-3}$)	0.93 %	$4.93*10^{-3}$ ($1.47*10^{-3}$)	1.79 %
IC, Q=10	$0.50*10^{-3}$ ($0.35*10^{-3}$)	5.32 %	$-0.75*10^{-3}$ ($1.19*10^{-3}$)	-0.72 %	$2.65*10^{-3}$ ($1.53*10^{-3}$)	0.96 %
IC, Q=20	$0.38*10^{-3}$ ($0.36*10^{-3}$)	4.03 %	$-0.43*10^{-3}$ ($1.15*10^{-3}$)	-0.41 %	$2.73*10^{-3}$ ($1.53*10^{-3}$)	0.99 %

Standard errors of bias estimates are below the estimates in parentheses.

Table 2: Simulation Estimates of Standard Errors of Estimators $\hat{F}_1.$, $\hat{F}_2.$ and $\hat{F}_3.$ under Correct Covariates and CME Nonresponse for DA-CME and MAR Nonresponse for DA-MAR and MAR-based Imputation.

Imputation Method	$se(\hat{F}_1.)$	$se(\hat{F}_2.)$	$se(\hat{F}_3.)$
DA-CME NN	$3.43*10^{-3}$	$11.35*10^{-3}$	$14.97*10^{-3}$
DA-MAR NN	$3.07*10^{-3}$	$10.76*10^{-3}$	$14.78*10^{-3}$
MAR-based NN	$3.07*10^{-3}$	$10.56*10^{-3}$	$14.57*10^{-3}$

Table 3: Simulation Estimates of Relative Biases and Coverage Rates of the 95% Confidence Intervals of $\hat{\text{var}}_{MI}(\hat{F}.)$ for the three Point Estimators using DA-CME under CME Nonresponse and Correct Covariates.

DA-CME	Rel. Bias $\hat{\text{var}}_{MI}(\hat{F}_1.)$	Rel. Bias $\hat{\text{var}}_{MI}(\hat{F}_2.)$	Rel. Bias $\hat{\text{var}}_{MI}(\hat{F}_3.)$	Coverage $\hat{\text{var}}_{MI}(\hat{F}_1.)$	Coverage $\hat{\text{var}}_{MI}(\hat{F}_2.)$	Coverage $\hat{\text{var}}_{MI}(\hat{F}_3.)$
Reg Imp	-1.27 %	-2.16 %	3.44 %	94 %	94 %	95 %
NN	-6.11 %	-3.45 %	6.15 %	93 %	95 %	96 %

Table 4: Simulation Estimates of Biases of Estimators $\hat{F}_1.$, $\hat{F}_2.$ and $\hat{F}_3.$ for Regression and Nearest Neighbour Imputation using Data Augmentation DA-CME under Misspecification of the Imputation Model, Assuming CME.

DA-CME	Bias of $\hat{F}_1.$	Rel. Bias of $\hat{F}_1.$	Bias of $\hat{F}_2.$	Rel. Bias of $\hat{F}_2.$	Bias of $\hat{F}_3.$	Rel. Bias of $\hat{F}_3.$
Reg Imp						
Misspecification 1	$0.31*10^{-3}$ ($0.21*10^{-3}$)	5.70 %	$-0.18*10^{-3}$ ($0.82*10^{-3}$)	-0.19 %	$-3.73*10^{-3}$ ($1.41*10^{-3}$)	-1.39 %
Misspecification 2	$0.39*10^{-3}$ ($0.21*10^{-3}$)	8.22 %	$-0.09*10^{-3}$ ($0.85*10^{-3}$)	-0.10 %	$-3.73*10^{-3}$ ($1.34*10^{-3}$)	-1.39 %
Misspecification 3	$0.32*10^{-3}$ ($0.19*10^{-3}$)	8.08 %	$1.78*10^{-3}$ ($0.80*10^{-3}$)	0.19 %	$-4.54*10^{-3}$ ($1.45*10^{-3}$)	-1.71%
$\ln(x_i)$ not generated	$0.09*10^{-3}$ ($0.22*10^{-3}$)	2.49 %	$-0.09*10^{-3}$ ($0.61*10^{-3}$)	-0.17 %	$0.48*10^{-3}$ ($1.34*10^{-3}$)	0.20 %
NN						
Misspecification 1	$-0.01*10^{-3}$ ($0.25*10^{-3}$)	-0.16 %	$1.25*10^{-3}$ ($1.07*10^{-3}$)	1.34 %	$6.08*10^{-3}$ ($1.78*10^{-3}$)	2.28 %
Misspecification 2	$0.12*10^{-3}$ ($0.25*10^{-3}$)	2.54 %	$2.43*10^{-3}$ ($1.02*10^{-3}$)	2.59 %	$7.28*10^{-3}$ ($1.80*10^{-3}$)	2.74 %
Misspecification 3	$0.06*10^{-3}$ ($0.22*10^{-3}$)	1.56 %	$3.48*10^{-3}$ ($1.04*10^{-3}$)	3.71 %	$7.39*10^{-3}$ ($1.71*10^{-3}$)	2.78 %
$\ln(x_i)$ not generated	$-0.07*10^{-3}$ ($0.21*10^{-3}$)	-2.01 %	$-0.81*10^{-3}$ ($0.75*10^{-3}$)	-1.61 %	$-1.39*10^{-3}$ ($1.55*10^{-3}$)	-0.59 %

Table 5: Simulation Estimates of Biases of Estimators $\hat{F}_1.$, $\hat{F}_2.$ and $\hat{F}_3.$ for Regression and Nearest Neighbour Imputation using Data Augmentation (DA-CME) under Misspecification of the Nonresponse Mechanism, Assuming a Correct Imputation Model.

DA-CME	Bias of $\hat{F}_1.$	Rel. Bias of $\hat{F}_1.$	Bias of $\hat{F}_2.$	Rel. Bias of $\hat{F}_2.$	Bias of $\hat{F}_3.$	Rel. Bias of $\hat{F}_3.$
Reg Imp						
CME+6	$0.14*10^{-3}$ ($0.28*10^{-3}$)	1.47 %	$0.17*10^{-3}$ ($0.91*10^{-3}$)	0.16 %	$0.33*10^{-3}$ ($1.25*10^{-3}$)	0.12 %
full model	$-0.43*10^{-3}$ ($0.28*10^{-3}$)	-4.61 %	$-6.83*10^{-3}$ ($0.96*10^{-3}$)	-6.56 %	$-16.6*10^{-3}$ ($1.44*10^{-3}$)	-6.07 %
MAR	$-0.21*10^{-3}$ ($0.29*10^{-3}$)	-2.20 %	$-7.14*10^{-3}$ ($0.93*10^{-3}$)	-6.86 %	$-23.19*10^{-3}$ ($1.20*10^{-3}$)	-8.46 %
NN						
CME+6	$-1.19*10^{-3}$ ($0.32*10^{-3}$)	-1.25 %	$1.64*10^{-3}$ ($1.14*10^{-3}$)	1.58 %	$5.69*10^{-3}$ ($1.48*10^{-3}$)	2.07 %
full model	$-2.49*10^{-3}$ ($3.26*10^{-3}$)	-2.61 %	$0.45*10^{-3}$ ($1.37*10^{-3}$)	0.43 %	$7.40*10^{-3}$ ($2.01*10^{-3}$)	2.70 %
MAR	$-0.35*10^{-3}$ ($0.32*10^{-3}$)	-3.77 %	$-1.69*10^{-3}$ ($1.13*10^{-3}$)	-1.63 %	$-4.54*10^{-3}$ ($1.45*10^{-3}$)	-1.65 %

Table 6: Simulation Estimates of Biases of Estimators $\hat{F}_1.$, $\hat{F}_2.$ and $\hat{F}_3.$ for MAR-based Regression and Nearest Neighbour Imputation under a CME Nonresponse Mechanism.

MAR-based	Bias of $\hat{F}_1.$	Rel. Bias of $\hat{F}_1.$	Bias of $\hat{F}_2.$	Rel. Bias of $\hat{F}_2.$	Bias of $\hat{F}_3.$	Rel. Bias of $\hat{F}_3.$
Reg Imp	$0.43*10^{-3}$ ($0.29*10^{-3}$)	4.61 %	$6.86*10^{-3}$ ($0.97*10^{-3}$)	6.59 %	$15.95*10^{-3}$ ($1.35*10^{-3}$)	5.81 %
NN	$0.30*10^{-3}$ ($0.32*10^{-3}$)	3.17 %	$6.77*10^{-3}$ ($1.14*10^{-3}$)	6.51 %	$16.57*10^{-3}$ ($1.62*10^{-3}$)	6.04 %

Table 7: Estimates $\hat{F}_1.$, $\hat{F}_2.$ and $\hat{F}_3.$ for 18+ (unweighted) using DA-CME, DA-MAR and other MAR-based Imputation Methods using Random Regression, Nearest Neighbour and Hot Deck Imputation Within Classes, Applied to March-May 2000.

	$\hat{F}_1.$ in %	$\hat{F}_2.$ in %	$\hat{F}_3.$ in %
DA-CME			
Reg Imp	0.55	1.91	22.39
NN	0.45	2.17	26.78
IC	0.44	2.35	26.35
DA-MAR			
Reg Imp	1.22	1.93	24.90
NN	0.50	2.29	28.78
IC	0.51	2.49	26.99
MAR-based			
Reg Imp	1.24	1.89	26.27
NN	0.50	2.27	29.04
IC	0.50	2.43	29.11