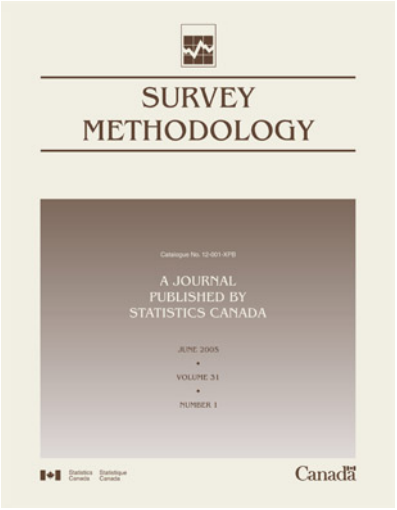
Catalogue no. 12-001-XIE

Survey Methodology

June 2006



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

June 2006

Published by authority of the Minister responsible for Statistics Canada

July 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 32, No. 1, pp. 25-36
Statistics Canada, Catalogue No. 12-001

Using Missing Data Methods to Correct for Measurement Error in a Distribution Function

Gabriele B. Durrant and Chris Skinner [1]

Abstract

This paper considers the use of imputation and weighting to correct for measurement error in the estimation of a distribution function. The paper is motivated by the problem of estimating the distribution of hourly pay in the United Kingdom, using data from the Labour Force Survey. Errors in measurement lead to bias and the aim is to use auxiliary data, measured accurately for a subsample, to correct for this bias. Alternative point estimators are considered, based upon a variety of imputation and weighting approaches, including fractional imputation, nearest neighbour imputation, predictive mean matching and propensity score weighting. Properties of these point estimators are then compared both theoretically and by simulation. A fractional predictive mean matching imputation approach is advocated. It performs similarly to propensity score weighting, but displays slight advantages of robustness and efficiency.

Key Words: Donor imputation; Fractional imputation; Hot deck imputation; Multiple imputation; Nearest neighbour imputation; Predictive mean matching; Propensity score weighting.

1. Introduction

Measurement error may lead to biased estimation of distribution functions (Fuller 1995). In this paper we consider approaches to correcting for this bias when, in addition to sample observations on the erroneously measured variable, values of the accurately measured variable are available for a subsample. When the subsample is selected using a randomised scheme, the set-up is an instance of the well-studied problem of double sampling (*e.g.*, Tenenbein 1970). In this case, unbiased estimates can be constructed from the subsample alone, but use of data on the correlated surrogate variable for the whole sample may improve efficiency. See, for example, Luo, Stokes and Sager (1998). In this paper we shall suppose that the subsample is not selected by a known randomised scheme, but rather by an unknown missing data mechanism. We shall just assume that the accurate variable is missing at random (MAR) (Little and Rubin 2002), conditional on variables measured on the whole sample. Some inference methods are available for this problem if we are willing to make strong parametric assumptions about the true distribution (*e.g.*, Buonaccorsi 1990) or about the measurement error model (*e.g.*, Luo *et al.* 1998). We shall not consider such methods further, however, since we suppose that we are dealing with an application where such assumptions are unrealistic. Instead, the novel feature of this paper is to view inference in this measurement error set-up as a missing data problem and to consider the application of imputation and weighting methods from the missing data literature. Our focus will be on the choice of such methods to improve point estimation of the distribution function, in terms of bias, efficiency and robustness to model assumptions. We shall only consider variance estimation briefly.

This paper is motivated by an application to the estimation of the distribution of hourly pay in the United Kingdom (UK), using data from the UK Labour Force Survey (LFS). In the LFS there are two ways of measuring hourly pay. The traditional method is to obtain information about earnings and hours worked and to derive a measure of hourly pay from this information. We refer to the variable derived in this way as the *derived hourly pay* variable. A more recent method of measuring hourly pay is to ask respondents directly about their hourly pay. We refer to the resulting measure of hourly pay as the *direct variable*. Skinner, Stuttard, Beissel-Durrant and Jenkins (2002) describe and provide empirical evidence of many sources of measurement error in the derived variable and conclude from their study that the direct variable measures hourly pay much more accurately than the derived variable. The problem with the direct variable is that it is missing for about 43% of all cases. The application is outlined in Section 8 and described in greater detail in Skinner *et al.* (2002), who also proposed the use of imputation to address the measurement error problem. This paper extends that work by considering a wider class of approaches to missing data and by comparing their properties both theoretically and via simulation. The imputation approach developed in this paper, which extends that considered by Skinner *et al.* (2002), has now been implemented by the UK Office for National Statistics as a new approach to producing low pay estimates.

The paper is structured as follows. The estimation problem is discussed in section 2. Imputation and weighting

1. Gabriele B. Durrant and Chris Skinner, University of Southampton, United Kingdom. E-mail: cjs@soton.ac.uk.

approaches are set out in sections 3 and 4 respectively and their properties are studied and compared theoretically in section 5 and via a simulation study in section 7. Variance estimation is considered briefly in section 6. Section 8 discusses the application of the methods to the LFS. Some concluding remarks are given in section 9.

2. The Estimation Problem

Let y_i be the (true) value of a variable of interest associated with unit i in a finite population U. The distribution function of the variable in U is:

$$F(y) = N^{-1} \sum_{i \in U} I(y_i \leq y), \tag{1}$$

where $I(.)$ is the truth function ($I(E) = 1$ if E is true and $= 0$ otherwise) and y may take any specified value. Suppose that a survey is conducted on a sample $s \subset U$ and that the variable is measured as y_i^* for units $i \in s$. The difference between y_i^* and y_i represents measurement error. Suppose that the true value y_i is recorded for a subset of sample units and that we write $r_i = 1$ if y_i is recorded and $r_i = 0$ otherwise. Let x_i be a vector of auxiliary variables also recorded in the survey. Our data consist of values y_i^*, x_i and r_i for $i \in s$ and values y_i for $i \in s$ when $r_i = 1$. The problem is how to use these data to make inference about $F(y)$.

In the LFS application, the units are employees, s is the set of unit respondents in the LFS sample, y_i^* is the value of the derived hourly pay variable and y_i is the value of the direct variable for employee i. The value y_i is assumed equal to the true hourly pay.

The primary feature of this inference problem that concerns us is the missingness of y_i values and we consider two approaches to handle this missingness:

- imputation of y_i for units $i \in s$ where $r_i = 0$, using the values y_i^* and x_i as auxiliary information;
- weighting of an estimator based upon the responding subsample $s_1 = \{i \in s;\ r_i = 1\}$, in particular, the use of propensity score weighting (Little 1986).

These approaches to estimating $F(y)$ will be discussed in the following two sections.

Inference will be discussed under a model-based framework, in which it is assumed that the population values (y_i, y_i^*, x_i, r_i), $i \in U$, are independently and identically (IID) distributed and that sampling is ignorable, that is the distribution of (y_i, y_i^*, x_i, r_i) is the same whether or not $i \in s$. In section 8 we shall comment on how the methods developed under these assumptions may be adapted to handle the sampling design of the LFS and the use of weights to compensate for unit non-response in the survey.

3. Imputation Approaches

Suppose initially that it is possible to observe y_i for all $i \in s$. Then, under the assumptions given in the previous section,

$$\hat{F}(y) = n^{-1} \sum_{i=1}^{n} I(y_i < y) \tag{2}$$

would be an unbiased estimator of $F(y)$, in the sense that $E[\hat{F}(y) - F(y)] = 0$ for all y, where we write $s = \{1,...,n\}$ and the expectation is with respect to the model, conditional on the selected sample s. To address the problem that y_i is missing when $r_i = 0$, suppose that y_i is replaced in (2) by an imputed value y_i^I when $r_i = 0$ (and $i \in s$) and let $\tilde{y}_i = y_i$ if $r_i = 1$ and $\tilde{y}_i = y_i^I$ otherwise. The resulting estimator of $F(y)$ is

$$\tilde{F}(y) = n^{-1} \sum_{i=1}^{n} I(\tilde{y}_i < y). \tag{3}$$

A sufficient condition for $\tilde{F}(y)$ to be an unbiased estimator of $F(y)$ is that the conditional distribution of y_i^I given $r_i = 0$, denoted $f(y_i^I \mid r_i = 0)$, is the same as the conditional distribution $f(y_i \mid r_i = 0)$. However, since y_i is only observed when $r_i = 1$, the data provide no direct information about $f(y_i \mid r_i = 0)$ without further assumptions. We consider two possible assumptions.

Assumption (MAR): r_i and y_i are conditionally independent given y_i^* and x_i.

Assumption (Common Measurement Error Model): r_i and y_i^* are conditionally independent given y_i and x_i.

The first assumption is the standard one made when using imputation or weighting (Little and Rubin 2002) and is the one which we shall make. The second assumption is that the measurement error model, defined as the conditional distribution of y_i^* given y_i and x_i, is the same for respondents ($r_i = 1$) and nonrespondents ($r_i = 0$). We shall use the second assumption in the simulation study in section 7 to assess robustness of MAR-based procedures. Inference under the second assumption is more difficult, however, and appears to require stronger modelling assumptions about the distribution of y_i and x_i; we are considering this problem in other research and do not pursue this further in this paper. The plausibility of these two assumptions for the LFS application is discussed further in Skinner *et al.* (2002).

Under the MAR assumption we have $f(y_i \mid y_i^*,\ x_i,\ r_i = 0) = f(y_i \mid y_i^*,\ x_i,\ r_i = 1)$ and a sufficient condition for $\tilde{F}(Y)$ to estimate $F(Y)$ unbiasedly is that

$$f(y_i^I \mid y_i^*,\ x_i,\ r_i = 0) = f(y_i \mid y_i^*,\ x_i,\ r_i = 1). \tag{4}$$

We therefore consider an imputation approach where the conditional distribution of y given y^* and x is 'fitted' to the respondent $(r_i = 1)$ data and then the imputed values y_i^I are 'drawn from' this fitted distribution at the values y_i^* and x_i observed for the nonrespondents. Suppose that the conditional distribution $f(y_i \mid y_i^*, x_i, r_i = 1)$ may be represented by a parametric regression model:

$$g(y_i) = h(y_i^*, x_i; \beta) + e_i, \; E(e_i \mid y_i^*, x_i) = 0 \qquad (5)$$

where $g(.)$ and $h(.)$ are given functions and β is a vector of regression parameters. A point predictor of y_i, given an estimator $\hat{\beta}$ of β based on respondent data, is

$$\hat{y}_i = g^{-1}[h(y_i^*, x_i; \hat{\beta})]. \qquad (6)$$

Using $\hat{y}_i$ for imputation may, however, lead to serious underestimation of $F(y)$ for low values of y, since such simple regression imputation is expected to reduce the variation in $F(y)$ artificially (Little and Rubin 2002, page 64). This effect might be avoided by taking $y_i^I = g^{-1}[h(y_i^*, x_i; \hat{\beta}) + \hat{e}_i]$, where $\hat{e}_i$ is a randomly selected empirical residual (Little and Rubin 2002, page 65). Our experience is, however, that this approach fails to generate imputed values which reproduce the 'spiky' behaviour of hourly pay distributions in our application and may lead to bias around these spikes. We prefer therefore to restrict attention to donor imputation methods, which set $y_i^I = y_{d(i)}$ $(r_i = 0)$ for some donor respondent $j = d(i)$ for which $r_j = 1$. The imputed value from a donor will always be a genuine value and will respect the spiky behaviour in our application. The basic donor imputation method we consider is predictive mean matching (Little 1988), that is nearest neighbour imputation with respect to $\hat{y}_i$, defined by (6), *i.e.*,

$$\text{impute } y_i \text{ by } y_{d(i)}$$
$$\text{satisfying } |\hat{y}_i - \hat{y}_{d(i)}| = \min_{j:\, r_j = 1} |\hat{y}_i - \hat{y}_j| \qquad (7)$$

where $r_i = 0$ and $r_{d(i)} = 1$.

Corollary 2 of Theorem 1 of Chen and Shao (2000) then provides theoretical justification for the approximate unbiasedness of the resulting estimator $\tilde{F}(y)$ for $F(y)$, if the following four conditions hold: (i) y_i is missing at random (MAR) conditional on $z_i = g^{-1}[h(y_i^*, x_i; \beta)]$, where $\beta = \text{plim}(\hat{\beta})$, (ii) the conditional expectation of y_i given z_i is monotonic and continuous in z_i, (iii) z_i and $E(y_i \mid z_i)$ have finite third moments and (iv) the probability of response given z is bounded above zero. These conditions seem plausible provided: the MAR assumption above holds; the distribution of y_i only depends on y_i^* and x_i via z_i; y_i^* is a reasonably good proxy for y_i. In addition, Chen and Shao's (2000) result needs to be adapted for the fact that the nearest neighbour is defined with respect to $\hat{\beta}$ whereas the above conditions are with respect to β. This adaptation seems plausible since, for a sufficiently large number of respondents, close neighbours with respect to $\hat{y}_i = g^{-1}[h(y_i^*, x_i; \hat{\beta})]$ should also be close neighbours with respect to $z_i = g^{-1}[h(y_i^*, x_i; \beta)]$.

There are thus theoretical grounds that nearest neighbour imputation with respect to $\hat{y}_i$ will lead to an approximately unbiased estimator of $F(y)$, subject to the MAR assumption and certain additional plausible conditions. It is also of interest to consider the efficiency of $\tilde{F}(y)$. The variance of $\tilde{F}(y)$ for nearest neighbour imputation may be inflated if certain donors may be used much more frequently than others. We consider a number of approaches to reducing this variance inflation effect.

First, we may restrict the number of times that respondents are used as donors by defining imputation classes by disjoint intervals of values of $\hat{y}_i$ and drawing donors for a recipient by simple random sampling from the class within which the recipient's value $\hat{y}_i$ falls. The smoothing will be greatest if we draw donors without replacement. We denote this hot deck method HDIWR or HDIWOR, depending on whether sampling is with or without replacement. A second approach is to undertake donor selection sequentially and to penalize the distance function employed for determining the nearest neighbour $d(i)$ as follows

$$|\hat{y}_i - \hat{y}_{d(i)}|(1 + \mu t_{d(i)}) = \min_{j:\, r_j = 1} \{|\hat{y}_i - \hat{y}_j|(1 + \mu t_j)\}, \qquad (8)$$

where $\mu \in \mathbb{R}^+$ is a penalty factor, t_j is the number of times the respondent j has already been used as a donor, $r_i = 0$ and $r_{d(i)} = 1$ (Kalton 1983). A third approach is to employ repeated imputed values $y_i^{I(m)}$, $m = 1, ..., M$, for each recipient $i \in s$ such that $r_i = 0$. The resulting estimator of $F(y)$ is $M^{-1}\sum_m \tilde{F}^{(m)}(y)$, the mean of the resulting estimators $\tilde{F}^{(m)}(y)$. We refer to the third approach as fractional imputation (Kalton and Kish 1984; Fay 1996) rather than multiple imputation (Rubin 1996), since we do not require the imputation method to be 'proper', that is to fulfil conditions which ensure that the multiple imputation variance estimator is consistent. We do not stipulate this requirement here because our primary objective is point estimation. In our use of fractional imputation we aim to select donors $d(i, m)$, $m = 1, ..., M$, each a close neighbour to i, so that $\tilde{F}^{(m)}(y)$ remains approximately unbiased for $F(y)$. We consider the following variations of this approach.

(i) The $M/2$ nearest neighbours above and below $\hat{y}_i$ are taken, for $M = 2$ or 10, denoted NN2 and NN10 respectively.

(ii) $M/2$ donors are selected by simple random sampling with replacement from the M respondents above and from the M respondents below $\hat{y}_i$, for $M = 2$ or 10, denoted NN2(4) and NN10(20) respectively.

(iii) $M = 10$ donors are selected by simple random sampling with or without replacement from the imputation classes referred to in the HDIWR and HDIWOR methods described above. We refer to these as the HDIWR10 and HDIWOR10 methods.

For comparison we also consider the Approximate Bayesian Bootstrap method of multiple imputation (Rubin and Schenker 1986), denoted ABB10, defined with respect to the imputation classes referred to in the HDIWR and HDIWOR methods.

4. Weighted Estimation

The estimator $\tilde{F}(y)$ implied by the different imputation approaches considered in the previous section may be expressed in weighted form as:

$$\tilde{F}(y) = \sum_{i \in s_1} w_i I(y_i < y) / \sum_{i \in s_1} w_i, \qquad (9)$$

where $s_1 = \{i \in s; r_i = 1\}$ is the set of respondents and $w_i = 1 + d_i / M$, where d_i is the total number of times that respondent i is used as a donor over the M repeated imputations. Note that $\sum_{s_1} w_i = n$. Another choice of weight would be to set w_i equal to the reciprocal of an estimated value of the propensity score, $\Pr(r_i = 1 | y_i^*, x_i)$ (Little 1986). This approach has been proposed for the hourly pay application by Dickens and Manning (2004). The propensity score might be estimated, for example, under a logistic regression model relating r_i to y_i^* and x_i. Under the MAR assumption, the resulting estimator $\tilde{F}(y)$ will be approximately unbiased assuming validity of the model for the conditional distribution $f(r_i | y_i^*, x_i)$ and some regularity conditions, such as those described in section 3 for the imputed estimator. Note that the need to model $f(r_i | y_i^*, x_i)$ replaces the need to model $f(y_i | y_i^*, x_i)$ in the imputation approach.

5. Properties of Imputation and Weighting Approaches

In this section we investigate and compare the theoretical properties of the imputation and propensity score weighting approaches introduced in the previous two sections under various simplifying assumptions. We fix y and set $u_i = I(y_i < y)$. Letting $N \to \infty$ we suppose that the parameter of interest is $\theta = E(u_i)$. We consider the imputation approach first and suppose that y_i depends upon y_i^* and x_i only via $z_i = g^{-1}[h(y_i^*, x_i; \beta]$ and that y_i is missing at random given z_i. Ignoring the difference between β and $\hat{\beta}$, assuming s_1 is large, we consider nearest neighbour imputation with respect to z_i. As in (9) the imputed estimator of θ may be expressed as

$$\hat{\theta}_{\mathrm{IMP}} = \sum_{i \in s_1} w_i u_i / \sum_{i \in s_1} w_i \qquad (10)$$

where $w_i = 1 + d_i / M$ (and $\sum_{s_1} w_i = n$). We write the corresponding expression for propensity score weighting as $\hat{\theta}_{\mathrm{PS}}$ with w_i replaced by $w_{\mathrm{PS}i}$. Let $z_{\mathrm{PS}i}$ be the scalar function of y_i^*, x_i upon which r_i depends and write:

$$\Pr(r_i = 1 | y_i^*, x_i) = \pi(z_{\mathrm{PS}i}). \qquad (11)$$

Just as we ignored the difference between β and $\hat{\beta}$, we initially ignore error in estimating $\pi(z_{\mathrm{PS}i})$ and write $w_{\mathrm{PS}i} = \pi(z_{\mathrm{PS}i})^{-1}$.

The imputation and propensity score weighting approaches may be expected to yield similar estimators if z_i and $z_{\mathrm{PS}i}$ are similar, that is they are close to deterministic functions of each other, and M is large. To see this, consider a simple example of the imputation approach, where the donor is drawn randomly from an imputation class c of close neighbours with respect to z_i, containing m_c respondents and $n_c - m_c$ nonrespondents, as described in section 3. In this case, w_i will approach $1 + (n_c - m_c)/m_c = n_c / m_c$ as $M \to \infty$ and this is the inverse of the response rate within the class (David, Little, Samuhel and Triest 1983). More generally, with the fractional nearest neighbour imputation approach considered in section 3, the weight $w_i = 1 + d_i / M$ may be interpreted as a local (with respect to z_i) nonparametric estimate of $\Pr(r_i = 1 | z_i)^{-1}$ despite the fact that imputation is based upon a model for y_i given z_i rather than r_i given z_i. Thus, the imputation approach may be expected to lead to similar estimation results to propensity score weighting if z_i and $z_{\mathrm{PS}i}$ are deterministic functions of each other. In general, however, this will not be the case. Since $\Pr(r_i = 1 | z_i)$ may be expressed as an average of $\Pr(r_i = 1 | y^*, x)$ across values of y^* and x for which $z = z_i$, we may interpret w_i as a smoothed version of $w_{\mathrm{PS}i}$ and may expect it to show less dispersion. This suggests that it may be possible to use imputation to improve upon the efficiency of estimates based on propensity score weighting, as also discussed by David *et al.* (1983) and Rubin (1996, section 4.6). To investigate this further, assuming MAR and the other assumptions in sections 3 and 4 upon which the approaches are based, both imputation and weighting approaches lead to approximately unbiased estimation of $F(y)$ and we may focus our comparison on relative efficiency.

It follows from equation (3.3) of Chen and Shao (2000) that the variance of $\hat{\theta}_{\mathrm{IMP}}$ may be approximated for large n by

$$\mathrm{var}(\hat{\theta}_{\mathrm{IMP}}) \approx n^{-2} E\left[\sum_{s_1} w_i^2 V(u_i | z_i)\right] + n^{-1} V[\psi(z_i)], \qquad (12)$$

where $\psi(z_i) = E(u_i \mid z_i)$ and any impact of estimating β is ignored. Note that Chen and Shao (2000) consider single imputation with $M = 1$ but their proof of this result carries through if $M > 1$. It is convenient to reexpress this result as

$$\text{var}(\hat{\theta}_{\text{IMP}}) \approx n^{-1}\sigma^2 + n^{-2}E\left[\sum_{s_1}(w_i^2 - w_i)V(u_i \mid z_i)\right], \quad (13)$$

using the identity

$$V[\psi(z_i)] = \sigma^2 - E[V(u_i \mid z_i)], \quad (14)$$

where $\sigma^2 = V(u_i)$ and a corollary of Chen and Shao's (2000) Theorem 1 that

$$E\left[n^{-1}\sum_{s_1} w_i V(u_i \mid z_i)\right] = E[V(u_i \mid z_i)] + o_p(n^{-1/2}). \quad (15)$$

Note that $w_i^2 - w_i = (d_i / M)(1 + d_i / M) \geq 0$. Expression (13) may be interpreted from both 'missing data' and 'measurement error' perspectives. From a missing data perspective, the first term in (13) is just the variance of $\hat{\theta}$ in the absence of missing data and the second term represents the inflation of this variance due to imputation error. From a measurement error perspective, we may consider limiting properties under 'small measurement error asymptotics' (Chesher 1991), that is where $y_i^* \to y_i$ and $V(u_i \mid z_i)$ approaches zero. In this case, the second term also approaches zero and $\hat{\theta}_{\text{IMP}}$ becomes 'fully efficient', *i.e.*, its variance approaches σ^2 / n.

Let us now consider propensity score weighting. We make the corresponding assumption that y_i is missing at random given $z_{\text{PS}i}$. Linearising the ratio in (9), with $w_{\text{PS}i}$ in place of w_i, using the fact that $E(\sum_{s_1} w_{\text{PS}i}) = n$ and initially ignoring the impact of estimating the propensity score we may write

$$\begin{aligned}\text{var}(\hat{\theta}_{\text{PS}}) &\approx n^{-2}\,\text{var}\left[\sum_{s_1} w_{\text{PS}i}(u_i - \theta)\right] \\ &= n^{-1}\,E[w_{\text{PS}i}(u_i - \theta)^2], \end{aligned} \quad (16)$$

which may be expressed alternatively as

$$\begin{aligned}\text{var}(\hat{\theta}_{\text{PS}}) \approx\ & n^{-2}E\left[\sum_{s_1} w_{\text{PS}i}^2 V(u_i \mid z_{\text{PS}i})\right] \\ & + n^{-1}E\{w_{\text{PS}i}[\psi(z_{\text{PS}i}) - \theta]^2\} \end{aligned} \quad (17)$$

To compare the efficiency of weighting and imputation it is convenient to use (14) and (15) (which hold also with $w_{\text{PS}i}$ in place of w_i) to obtain

$$\begin{aligned}\text{var}(\hat{\theta}_{\text{PS}}) \approx\ & n^{-1}\sigma^2 \\ & + n^{-2}E\left[\sum_{s_1}(w_{\text{PS}i}^2 - w_{\text{PS}i})V(u_i \mid z_{\text{PS}i})\right] \\ & + n^{-1}E\left\{\sum_{s_1}[w_{\text{PS}i} - 1][\psi(z_{\text{PS}i}) - \theta]^2\right\}. \end{aligned} \quad (18)$$

Note that, in comparison with (13), this involves a third term, which does not necessarily converge to zero as y_i^* approaches y_i and $V(u_i \mid z_{\text{PS}i}) \to 0$. Hence propensity score weighting does not become fully efficient as the measurement error disappears. The second term of (18) may also be expected to dominate the second term of (13) when $V(u_i \mid z_i)$ and $V(u_i \mid z_{\text{PS}i})$ are constant and equal, since, recalling that $\sum_{s_1} w_i = E(\sum_{s_1} w_{\text{PS}i}) = n$, these second terms are primarily determined by the variances of the weights w_i and $w_{\text{PS}i}$, and, provided M is sufficiently large, we may expect w_i to display less variation than $w_{\text{PS}i}$, as argued above.

The above discussion ignores the potential impact of estimating β or estimating a parameter vector α upon which the propensity score $\Pr(r_i = 1 \mid y_i^*, x_i)$ may be assumed to depend. Kim (2004) shows in fact that the estimation of α by its maximum likelihood estimator $\hat{\alpha}$ reduces the variance of $\hat{\theta}_{\text{PS}}$ as follows:

$$\begin{aligned}\text{var}(\hat{\theta}_{\text{PS}}) \approx\ & \text{var}(\tilde{\theta}_{\text{PS}}) \\ & - \text{cov}(\tilde{\theta}_{\text{PS}}, \hat{\alpha})\,\text{var}(\hat{\alpha})^{-1}\,\text{cov}(\hat{\alpha}, \tilde{\theta}_{\text{PS}}), \end{aligned} \quad (19)$$

where $\tilde{\theta}_{\text{PS}}$ is the estimator $\hat{\theta}_{\text{PS}}$ with the estimated propensity scores replaced by their true values and where the left hand sides of (16), (17) and (18) should now be $\text{var}(\tilde{\theta}_{\text{PS}})$. We conclude from this fact and the previous discussion that, in general, $\hat{\theta}_{\text{IMP}}$ is not necessarily more efficient than $\hat{\theta}_{\text{PS}}$ or vice versa and we look to the simulation study in section 7 for numerical evidence. However, our conclusion that $\hat{\theta}_{\text{IMP}}$ is more efficient as measurement error disappears and $y_i^* \to y_i$ remains valid even in the presence of estimation error in α and β, since the impact of estimation error in β will disappear in this case with $z_i \to y_i^*$ whereas the second term in (19) when added to expression (18) will not in general reduce $\text{var}(\hat{\theta}_{\text{PS}})$ to σ^2 / n in this case.

Let us finally consider the impact of departures from the MAR assumption. Under small measurement error asymptotics where $y_i^* \to y_i$ and $V(u_i \mid z_i) \to 0$ so $y_i^I \to y_i$, the imputation approach will provide consistent inference about θ even if the MAR assumption fails. This is not the case for the propensity score weighting approach. This suggests that the imputation approach may display more robustness to departures from the MAR assumption if the amount of measurement error is relatively small.

6. Variance Estimation

Although point estimation is the primary focus of this paper, we do now consider linearization variance estimation briefly. For propensity score weighting we refer to Kim (2004). For the single and fractional imputation methods in section 3 based upon nearest neighbour imputation, we may

consider a simplified approach based on the IID assumption set out in section 2 and the expression for the variance of $\hat{\theta}_{\text{IMP}}$ in (13).

The simple estimator of the first term σ^2/n:

$$n^{-1}\hat{\sigma}^2 = n^{-2}\sum_{s_1} w_i(u_i - \hat{\theta}_{\text{IMP}})^2 \qquad (20)$$

is approximately unbiased from Corollary 1 of Chen and Shao (2000). It follows that an approximately unbiased estimator of $\text{var}(\hat{\theta}_{\text{IMP}})$ is

$$\hat{V}(\hat{\theta}_{\text{IMP}}) = n^{-1}\hat{\sigma}^2 + n^{-2}\sum_{s_1}(w_i^2 - w_i)\hat{V}(u_i \mid \text{z}_i) \qquad (21)$$

if we can construct an approximately unbiased estimator $\hat{V}(u_i \mid \text{z}_i)$ of $V(u_i \mid \text{z}_i)$. Various approaches to estimating $V(u_i \mid \text{z}_i)$ seem possible. Following Fay (1999), we might consider the sample variance of u_j values for responding neighbours near to i with respect to z. An alternative approach would be to consider a model-based approach in which a model is fitted to $\psi(\text{z}_i) = E(u_i \mid \text{z}_i)$ for $i \in s$ giving $\hat{\psi}(\text{z}_i)$ and we set $\hat{V}(u_i \mid \text{z}_i) = \hat{\psi}(\text{z}_i)[1 - \hat{\psi}(\text{z}_i)]$. We have considered nonparametric methods of fitting $\psi(\text{z}_i)$, but have found with the LFS data that these lead to very similar values of $\hat{V}(\hat{\theta}_{\text{IMP}})$ as a logistic regression model for $\psi(\text{z}_i)$.

It may be possible to apply ideas in Chen and Shao (2001) or Kim and Fuller (2002) to extend the above approach to handle survey weights and a complex design. See Rancourt (1999) and Fay (1999) for other variance estimation approaches for nearest neighbour imputation and Little and Rubin (2002) for multiple imputation approaches.

7. Simulation Study

The aim of the study is to generate independent repeated samples $s^{(h)}$, $h = 1, \ldots, H$, with values y_i, y_i^*, x_i, r_i, $i \in s^{(h)}$ which are realistic in relation to the LFS application, considered further in section 8, to compute the corresponding estimates $\tilde{F}^{(h)}(y)$ for alternative approaches to missing data and values of y and to assess the performance of the estimators $\tilde{F}(y)$ empirically. In order to employ realistic values, the samples $s^{(h)}$ of size n were drawn with replacement (*i.e.*, using the bootstrap) from an actual sample of about 16,000 employees for the March–May 2000 quarter of the LFS (only main jobs of employees aged 18+ were considered and the very small number of cases with missing values on y_i^* or x_i were omitted). The values of x_i for each sample $s^{(h)}$ were taken directly from the values in the LFS sample. Variables were chosen for inclusion in x_i if they were either related to hourly pay, measurement error in y_i^* or response r_i (see Skinner *et al.* 2002) and included for example age, gender, household position, qualifications, occupation, duration of employment, full-time/part-time, industry and region (several of these variables were represented by dummy variables). We set $n = 15{,}000$, such that each $s^{(h)}$ was of a similar size as the original LFS sample, and $H = 1{,}000$. The values of y_i, y_i^* and r_i for each sample $s^{(h)}$ were generated from models, rather than directly from the LFS data, for the following reasons.

y_i: these values were generated from a model because they were frequently missing in the LFS. A linear regression model was used, relating $\ln(y_i)$ to $\ln(y_i^*)$ and x_i with a normal error and with 20 covariates including squared terms in $\ln(y_i^*)$ and age and interactions between $\ln(y_i^*)$ and 5 components of x_i. The model was fitted to the roughly 7,000 cases where y_i was observed.

y_i^*: these values were generated from a model to avoid duplicate values of (y_i^*, x_i) within each $s^{(h)}$, which it was considered might lead to an unrealistic distribution of distances between units for the nearest neighbour method. The model was a linear regression model relating $\ln(y_i^*)$ to x_i with a normal error and with 12 covariates, including a squared term in age and one interaction, fitted to the LFS data.

r_i: these values were generated from a model to ensure that the missing data mechanism was known. Several models were fitted. The only one reported here is a logistic regression relating r_i to $\ln(y_i^*)$ and x_i with 17 covariates including squared $\ln(y_i^*)$ and interactions between $\ln(y_i^*)$ and two covariates. The model was fitted to the LFS data. The missing data mechanism is MAR given the y_i^* and x_i for all the results presented except those in Table 5.

Estimates $\hat{\theta}_t^{(h)}$ of two parameters $(t = 1, 2)$ were obtained for each sample $s^{(h)}$,

$\theta_1 =$ proportion with pay below the national minimum wage (= £3.00 per hour aged $18-21$, £3.60 per hour aged 22+)

$\theta_2 =$ proportion with pay between minimum wage and £5/hour.

The true values are $\theta_1 = 0.056$ and $\theta_2 = 0.185$. The bias and standard error were estimated as $\hat{\text{bias}}(\hat{\theta}_t) = \bar{\theta}_t - \theta_t$ and $\hat{\text{s.e.}}(\hat{\theta}_t) = [H^{-1}\sum_{h=1}^{H}(\hat{\theta}_t^{(h)} - \bar{\theta}_t)^2]^{1/2}$, where $\bar{\theta}_t = H^{-1}\sum_h \theta_t^{(h)}$.

For the fractional imputation methods several different values for M were explored and $M = 10$ or 20 were chosen to achieve an increase in the efficiency whilst still being able to define a nearest neighbour imputation sensibly.

We first compare results for the alternative imputation approaches. Table 1 presents estimates of the biases of estimators of θ_1 and θ_2 for different imputation methods, for a MAR missing data mechanism. There is no evidence of significant biases for any of the nearest neighbour (NN) methods. The bias/standard error ratios are small and may be expected to be even smaller for estimates within domains *e.g.*, regions or age groups. We conclude that there is no evidence of important bias for these methods, provided the MAR mechanism holds and the model is correctly specified.

There is some evidence of statistically significant biases for each of the three methods based on imputation classes (HDIWR10, HDIWOR10, ABB10) perhaps because of the width of the classes, although the bias appears to be small relative to the standard error. Given the additional disadvantage of these methods, that the specification of the boundaries of the classes is arbitrary, these methods appear to be less attractive than the nearest neighbour methods. This finding contrasts with the preference sometimes expressed (*e.g.*, Brick and Kalton 1996, page 227) for stochastic methods of imputation, such as the HDI methods, compared to deterministic methods, such as nearest neighbour imputation, when estimating distributional parameters.

Corresponding estimates of standard errors are given in Table 2. We find as expected that the greatest standard error occurs for the single NN1 imputation method. The variance is reduced by around 10% using the penalty function method (NN1P). About $10-20\%$ reduction arises from using two imputations (NN2 or NN2 (4)) and around 20% reduction from using ten imputations (NN10, NN10 (20)), HDIWR10, HDIWOR10, ABB10). For a given number of imputations (2 or 10) there seem to be no obvious systematic effects of using a stochastic method (NN2 (4) or NN10 (20)) versus a deterministic method (NN2 or NN10). We would expect the standard errors for HDIWR10 to be no less than HDIWOR10, which is the case for $\hat{\theta}_1$ in table 2. The slight reduction for the standard error of estimator $\hat{\theta}_2$ is likely to be caused by a comparatively small number of simulation iterations $(H=1,000)$, which may not be fully sufficient for standard error estimation. We conclude that NN10 is the most promising approach, avoiding the bias of the imputation class methods and having appreciable efficiency gains over the methods generating one or two imputations.

Table 1
Simulation Estimates of Biases of Estimators of θ_1 and θ_2 for Different Imputation Methods, Assuming MAR and Correct Covariates $(H=1,000)$

Imputation Method	Bias of $\hat{\theta}_1$	Rel. Bias of $\hat{\theta}_1$	Bias of $\hat{\theta}_2$	Rel. Bias of $\hat{\theta}_2$
NN1	$1.2*10^{-4}$	0.2 %	$0.9*10^{-4}$	0.0 %
	$(0.9*10^{-4})$		$(1.7*10^{-4})$	
NN1P[1]	$4.4*10^{-4}$	0.8 %	$0.3*10^{-4}$	0.0 %
	$(2.6*10^{-4})$		$(5.1*10^{-4})$	
NN2	$0.6*10^{-4}$	0.1 %	$1.6*10^{-4}$	0.0 %
	$(0.8*10^{-4})$		$(1.5*10^{-4})$	
NN2(4)	$1.4*10^{-4}$	0.2 %	$-2.5*10^{-4}$	−0.1 %
	$(0.9*10^{-4})$		$(1.5*10^{-4})$	
NN10	$0.2*10^{-4}$	0.0 %	$-1.2*10^{-4}$	−0.1 %
	$(0.8*10^{-4})$		$(1.5*10^{-4})$	
NN10(20)	$0.2*10^{-4}$	0.0 %	$0.7*10^{-4}$	0.0 %
	$(0.8*10^{-4})$		$(1.5*10^{-4})$	
HDIWR10	$2.8*10^{-4}$	0.5 %	$26.2*10^{-4}$	1.4 %
	$(0.8*10^{-4})$		$(1.5*10^{-4})$	
HDIWOR10	$2.5*10^{-4}$	0.4 %	$28.0*10^{-4}$	1.5 %
	$(0.8*10^{-4})$		$(1.5*10^{-4})$	
ABB10	$4.6*10^{-4}$	0.8 %	$29.8*10^{-4}$	1.6 %
	$(0.8*10^{-4})$		$(1.5*10^{-4})$	

Standard errors of bias estimates are below the estimates in parentheses.
[1] Note: $H=100$ iterations were used due to computing time.

Table 2
Simulation Estimates of Standard Errors of Estimators of θ_1 and θ_2 for Different Imputation Methods, Assuming MAR and Correct Covariates ($H = 1,000$)

Imputation Method	s.e.$(\hat{\theta}_1)$	s.e.$(\hat{\theta}_2)$	$\frac{V(\hat{\theta}_1)}{V_{NN1}(\hat{\theta}_1)}$	$\frac{V(\hat{\theta}_2)}{V_{NN1}(\hat{\theta}_2)}$
NN1	$2.79*10^{-3}$	$5.43*10^{-3}$	1	1
NN1P[2]	$2.60*10^{-3}$	$5.15*10^{-3}$	0.87	0.91
NN2	$2.68*10^{-3}$	$5.05*10^{-3}$	0.91	0.86
NN2(4)	$2.73*10^{-3}$	$4.88*10^{-3}$	0.94	0.80
NN10	$2.56*10^{-3}$	$4.88*10^{-3}$	0.83	0.81
NN10(20)	$2.57*10^{-3}$	$4.79*10^{-3}$	0.84	0.77
HDIWR10	$2.52*10^{-3}$	$4.66*10^{-3}$	0.82	0.74
HDIWOR10	$2.48*10^{-3}$	$4.72*10^{-3}$	0.78	0.76
ABB10	$2.63*10^{-3}$	$4.87*10^{-3}$	0.88	0.80

[2] Note: $H = 100$ iterations were used due to computing time.

We next compare the NN10 imputation approach with propensity score weighting (PSW). We consider not only the case when the specification of the model used for imputation or weighting corresponds to the model used in the simulation, as in Table 1, but also some cases of misspecification. To ensure a fair comparison of weighting and imputation we use the same covariates when fitting both the models generating y_i and r_i. We first consider the estimated biases in Table 3. When the model for imputation (NN10) or the propensity scores is correctly specified neither method demonstrates any significant bias in the estimation of θ_1 or θ_2. Significant bias does arise, however, in both cases if the model is misspecified by failing to include covariates used in the simulation. The amount of bias is, however, noticeably greater for the weighting approach. For example, for the estimator $\hat{\theta}_1$ the bias is $3-7$ times higher under PSW than under NN10 depending on the misspecification. The impact of the misspecification seems higher for estimator $\hat{\theta}_2$, in particular for the PSW method. For this estimator, we found a $6-15$ times higher bias for PSW than for NN10.

Corresponding estimated standard errors of $\hat{\theta}_1$ and $\hat{\theta}_2$ are given in Table 4. These also tend to be greater for the weighting approach, showing an increase between $5-15\%$ in comparison to the imputation method. The increase in the standard error is higher for the second estimator $\hat{\theta}_2$, ranging from $12-15\%$, whereas for estimator $\hat{\theta}_1$ the increase is between $5-12\%$, depending on the misspecification. Consequently, the mean squared error is also higher for the weighting approach, with the increase ranging from 20% to 28% for the six values in Table 4. At least under the MAR assumption, the NN10 imputation approach appears to be preferable to propensity score weighting in terms of bias and variance.

Finally, we compare the properties of imputation (NN10) and propensity score weighting when the MAR assumption fails. We now simulate missingness according to the Common Measurement Error model assumption of section 3. The same logistic model with the same coefficients as in the previous simulation is used except that y_i^* is replaced as a covariate by y_i. Simulation estimates of biases and standard errors are presented in Table 5. We observe a non-negligible significant relative bias of around 5% for the imputation approach and a little higher for the propensity score weighting approach. The positive direction of the bias of $\hat{\theta}_1$ is as expected from arguments in Dickens and Manning (2004) and Skinner *et al.* (2002). MAR-based methods will tend to overestimate numbers of the low paid, if the CME assumption holds. This is because employees with observed y_i values tend to be lower paid than employees with missing y_i values and a MAR-based imputation method, even conditional on other variables, would tend to impute lower hourly pay values than would be the case under CME which allows for the dependency on true hourly pay. 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. The relative bias of 5% of the NN10 approach does not, however, appear to make the resulting estimates unusable.

Table 3
Simulation Estimates of Biases of Estimators of θ_1 and θ_2 for Nearest Neighbour Imputation (NN10) and Propensity Score Weighting, Assuming MAR and Correct and Misspecified Covariates ($H = 1{,}000$)

Method	Assumed Covariates	Bias of $\hat{\theta}_1$	Rel. Bias of $\hat{\theta}_1$	Bias of $\hat{\theta}_2$	Rel. Bias of $\hat{\theta}_2$
NN10	M1 (correct)	$-0.18*10^{-4}$	$-0.03\,\%$	$-5.8*10^{-4}$	$-0.31\,\%$
		$(0.64*10^{-4})$		$(1.20*10^{-4})$	
	M2	$-1.31*10^{-4}$	$-0.24\,\%$	$-4.74*10^{-4}$	$-0.25\,\%$
		$(0.65*10^{-4})$		$(1.23*10^{-4})$	
	M3	$-1.66*10^{-4}$	$-0.30\,\%$	$-10.6*10^{-4}$	$-0.57\,\%$
		$(0.63*10^{-4})$		$(1.23*10^{-4})$	
Propensity Score Weighting	M1 (correct)	$0.15*10^{-4}$	$0.03\,\%$	$-2.62*10^{-4}$	$-0.14\,\%$
		$(0.72*10^{-4})$		$(1.35*10^{-4})$	
	M2	$-8.96*10^{-4}$	$-1.64\,\%$	$70.2*10^{-4}$	$3.80\,\%$
		$(0.68*10^{-4})$		$(1.40*10^{-4})$	
	M3	$-5.02*10^{-4}$	$-0.92\,\%$	$67.8*10^{-4}$	$3.66\,\%$
		$(0.68*10^{-4})$		$(1.41*10^{-4})$	

Note: M1 is the correct model
M2 excludes the interactions and the square terms from the correct model
M3 drops further covariates from model M2.

Table 4
Simulation Estimates of Standard Errors of Estimators of θ_1 and θ_2 for Nearest Neighbour Imputation (NN10) and Propensity Score Weighting, Assuming MAR and Correct and Misspecified Covariates ($H = 1{,}000$)

Method	Assumed Covariates	s.e.($\hat{\theta}_1$)	s.e.($\hat{\theta}_2$)	MSE($\hat{\theta}_1$)	MSE($\hat{\theta}_2$)
NN10	M1 (correct)	$2.02*10^{-3}$	$3.80*10^{-3}$	$4.10*10^{-6}$	$1.49*10^{-5}$
	M2	$2.06*10^{-3}$	$3.88*10^{-3}$	$4.29*10^{-6}$	$1.54*10^{-5}$
	M3	$2.01*10^{-3}$	$3.89*10^{-3}$	$4.10*10^{-6}$	$1.63*10^{-5}$
Propensity Score Weighting	M1 (correct)	$2.27*10^{-3}$	$4.27*10^{-3}$	$5.16*10^{-6}$	$1.83*10^{-5}$
	M2	$2.17*10^{-3}$	$4.42*10^{-3}$	$5.51*10^{-6}$	$6.90*10^{-5}$
	M3	$2.16*10^{-3}$	$4.46*10^{-3}$	$4.94*10^{-6}$	$6.59*10^{-5}$

Table 5
Simulation Estimates of Biases and Standard Errors of Estimators of θ_1 and θ_2 for Nearest Neighbour Imputation (NN10) and Propensity Score Weighting. Under the (non-MAR) Common Measurement Error Model ($H = 1{,}000$)

Method	Bias of $\hat{\theta}_1$	Rel. Bias of $\hat{\theta}_1$	Bias of $\hat{\theta}_2$	Rel. Bias of $\hat{\theta}_2$	s.e.($\hat{\theta}_1$)	s.e.($\hat{\theta}_2$)
NN10	$29.0*10^{-4}$	$5.1\,\%$	$92.0*10^{-4}$	$5.0\,\%$	$2.53*10^{-3}$	$4.70*10^{-3}$
	$(0.8*10^{-4})$		$(1.48*10^{-4})$			
Propensity Score Weighting	$32.3*10^{-4}$	$5.7\,\%$	$100*10^{-4}$	$5.7\,\%$	$2.31*10^{-3}$	$4.42*10^{-3}$
	$(0.73*10^{-4})$		$(1.40*10^{-4})$			

8. Application to the Labour Force Survey

In this section we consider the application of the methods developed in sections $2-4$ to LFS data. The LFS provides an important source of estimates of the distribution of hourly pay in the UK (Stuttard and Jenkins 2001). It is a quarterly survey of households selected from a national file of postal addresses with equal probabilities by stratified systematic sampling. All adults in selected households are included in the sample. The resulting sample is clustered by

household membership but not by geography. Each selected household is retained in the sample for interview on five successive quarters and then rotated out and replaced. Questions relating to hourly pay are asked in just the first and fifth interviews, generating data on this topic for about 16,000 employees per quarter.

Two measures of hourly pay are constructed, as outlined in Section 1. The derived hourly pay variable in the LFS is defined as follows: (a) employees are asked questions about their main job to determine earnings over a reference period, (b) questions are asked to determine hours worked over the reference period and (c) the result of (a) is divided by the result of (b). The direct variable is obtained by first asking whether the respondent is paid a fixed hourly rate and then, if the answer is positive, by asking respondents what this (basic) rate is. Skinner *et al.* (2002) discuss how the derived variable suffers from many sources of measurement error, as in similar surveys in other countries (Rodgers, Brown and Duncan 1993; Moore, Stinson and Welniak 2000). They conclude that the direct variable measures hourly pay much more accurately. A working assumption in this application is that the direct variable measures hourly pay without error. The problem with the direct variable, however, is that it is missing for respondents who state that they are not paid at a fixed hourly rate (and for item nonrespondents) and this missingness is positively associated with hourly pay. The proportion of LFS respondents with a (main) job who provide a response to the direct question is about 43%. This proportion tends to be higher for lower paid employees, for example the rate is 72% among those in the bottom decile of the derived variable. The direct variable is not collected for second (and further) jobs and we therefore restrict attention only to main jobs. The aim is to use the missing data methods developed in this paper to correct for the measurement error in hourly pay. Skinner *et al.* (2002) discuss the plausibility of the two missing data assumptions in section 3 for this application.

The methods in sections $2-4$ were developed under the assumption of an IID model and ignorable sampling. Employees are selected with equal probabilities in the LFS so the sampling may be viewed as ignorable with respect to the bias of point estimation but unit non-response is likely to be differential and survey weights are constructed to compensate for this non-response (ONS 1999). We propose to incorporate these survey weights into the estimator in (3) or equivalently to multiply the weights w_i in (9) by the survey weights. This is analogous to the way the pseudo-likelihood approach (Skinner 1989) weights estimators based upon an IID assumption. The aim is to use the methods of sections $2-4$ to compensate for bias due to measurement error and item non-response and the survey weights to compensate for bias due to sampling and unit nonresponse. We have not attempted to take account of the weights in the imputation methods and this could be explored in future research.

We now apply nearest neighbour imputation, hot deck imputation within classes and propensity score weighting to LFS data. All methods are weighted by the survey weights. Figure 1 compares an estimated distribution, which ignores measurement error (the bold line) with estimates based on three missing data methods (the three dotted lines). We suggest the latter estimates are more approximately unbiased than the former estimate. All three missing data adjustments show, as expected, a strong 'kink' in the distribution at the level of the national minimum wage unlike for the derived variable. Corresponding estimates of two low pay proportions of interest are presented in Table 6. The 'missing data adjustments' have a substantial impact in comparison to estimates based on the derived variable. The results suggest that the proportion of jobs paid at or below the national minimum wage rate may be overestimated by four or five times if measurement error is ignored. The differences between the missing data methods are much smaller. We can see that the estimates under propensity score weighting differ from estimates derived using imputation methods, at least for the June-August 1999 quarter. Note that this quarter of the LFS was subject to a lower response rate than subsequent quarters resulting from changes in the LFS questionnaire. It was found that for consecutive quarters, which are subject to about 43% response rate, weighting and imputation led to very similar estimates of low pay proportions, as illustrated in table 7 for the March–May 2000 quarter. The decrease in the proportion of low paid employees over time is a result of the impact of the National Minimum Wage legislation. In addition, different imputation and propensity score models are used to analyse the effects of various model specifications on estimates of low pay. From Table 6 we can see that there is an indication that different models can have an effect on the estimates. With increasing complexity of the model a reduction in the estimates for both point estimators is observed. This might reflect a departure from the MAR assumption for the simpler imputation models. At least for the 1999 quarter, the differences in the estimates between weighting and imputation methods seem to be greater than between models. Note that the estimates presented here might differ slightly from official UK estimates since, for example, the official estimates are based on different imputation models, treating outliers differently or imputing differently for certain professions.

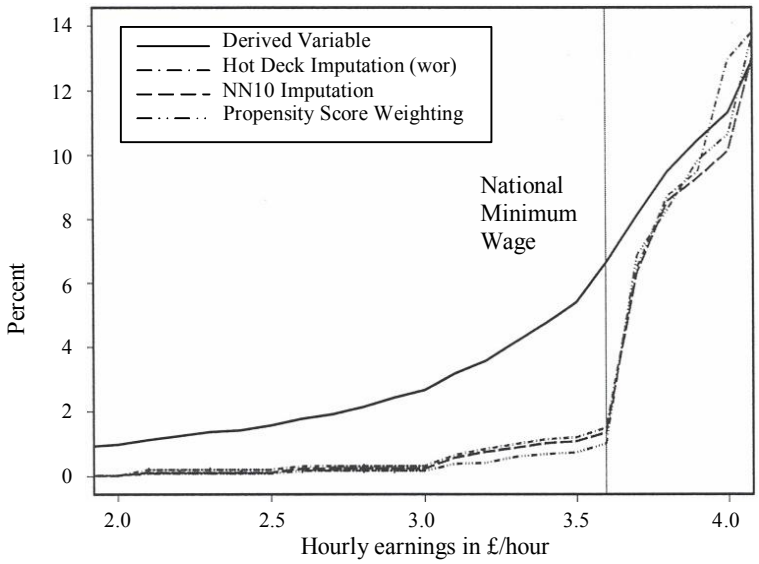


Figure 1. Alternative Estimates of the Distribution of Hourly Earnings From £2 to £4 for Age Group 22+, June-August 1999.

Table 6
Estimates of θ_1 and θ_2 (Weighted) for 18+ Using Different Propensity Score Models and Imputation Models Applied to LFS, June–August 1999

Method	Propensity Score Model or Imputation Model	(Weighted) $\hat{\theta}_1$ (%)	(Weighted) $\hat{\theta}_2$ (%)
Derived Variable	–	7.13	20.5
Propensity Score Weighting	M1	0.96	34.5
	M2	1.08	38.4
	M3	1.08	38.4
HDIWOR10	M1	1.44	32.1
	M2	1.41	32.9
	M3	1.50	33.2
NN10	M1	1.32	32.6
	M2	1.44	32.8
	M3	1.50	33.0

Note: M1 is the most complex model including square terms and interactions
M2 excludes the interactions and the square terms from model M1
M3 drops further covariates from model M2.

Table 7
Estimates of θ_1 and θ_2 (Weighted) for 18+ Using Propensity Score Weighting and Imputation Applied to LFS, March–May 2000

Method	Propensity Score Model or Imputation Model	(Weighted) $\hat{\theta}_1$ (%)	(Weighted) $\hat{\theta}_2$ (%)
Propensity Score Weighting	M1	0.54	27.10
HDIWOR10	M1	0.57	26.01
NN10	M1	0.55	26.61

9. Conclusions

In this paper we have considered the application of alternative missing data methods to correct for bias in the estimation of a distribution function arising from measurement error. Among imputation methods, nearest neighbour methods have performed most promisingly in terms of bias. These deterministic methods display no evidence of greater bias than stochastic imputation methods. Fractional imputation has shown appreciable efficiency gains compared to single imputation and appears more effective than penalizing the distance function or sampling without replacement with single imputation. In comparison to a propensity score weighting approach, the fractional nearest neighbour imputation has performed similarly, but has demonstrated slight advantages of robustness and efficiency. The simulation study suggested that the impact on the bias under a wrong model is greater for propensity score weighting and that the standard errors for the weighting approach were approximately $5-15\%$ times higher than for the imputation method.

Further research is being undertaken to develop and evaluate associated variance estimation methods, as well as alternative point estimation methods based upon the Common Measurement Error Model in section 2.

Acknowledgements

We are grateful to Danny Pfeffermann for comments on an earlier version of this paper.

References

Brick, J.M., and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

Buonaccorsi, J.P. (1990). Double sampling for exact values in some multivariate measurement error problems. *Journal of the American Statistical Association*, 85, 1075-1082.

Chen, J., and Shao, J. (2000). Nearest neighbour imputation for survey data. *Journal of Official Statistics*, 16, 113-131.

Chen, J., and Shao, J. (2001). Jackknife variance estimation for nearest neighbour imputation. *Journal of the American Statistical Association*, 96, 453, 260-269.

Chesher, A. (1991). The effect of measurement error. *Biometrika*, 78, 451-462.

David, M.H., Little, R., Samuhel, M. and Triest, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economic Statistics Section*, American Statistical Association, 168-173.

Dickens, R., and Manning, A. (2004). Has the national minimum wage reduced UK wage inequality? *Journal of the Royal Statistical Society*, Series A, 4, 613-626.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Fay, R.E. (1999). Theory and application of nearest neighbour imputation in census 2000. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 112-121.

Fuller, W.A. (1995). Estimation in the presence of measurement error. *International Statistical Review*, 63, 121-141.

Kalton, G. (1983). *Compensating for missing survey data*. Michigan, Institute for Social Research.

Kalton, G., and Kish, L. (1984). Some efficient random imputation methods. *Communications in Statistics*, *Part A*, *Theory and Methods*, 13, 1919-1939.

Kim, J.K. (2004). Efficient nonresponse weighting adjustment using estimated response probability. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Kim, J.-K., and Fuller, W.A. (2002). Variance estimation for nearest neighbour imputation. Unpublished manuscript.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.

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: John Wiley & Sons, Inc.

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.

Moore, J.C., Stinson, L.L. and Welniak, E.J. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics*, 16, 331-361.

ONS (1999). *Labour Force Survey*. User Guide, Volume 1, Background and Methodology, London.

Rancourt, E. (1999). Estimation with nearest neighbour imputation at Statistics Canada. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 131-138.

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. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

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

Skinner, C.J. (1989). Domain means, regression and multivariate analysis. In *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith), Chichester, Wiley.

Skinner, C., Stuttard, N., Beissel-Durrant, G. and Jenkins, J. (2002). The measurement of low pay in the UK Labour Force Survey. *Oxford Bulletin of Economics and Statistics*, 64, 653-676.

Stuttard, N., and Jenkins, J. (2001). Measuring low pay using the new earnings survey and the Labour Force Survey. *Labour Market Trends*, January 2001, 55-66.

Tenenbein, A. (1970). A double sampling scheme for estimating from binary data with misclassifications. *Journal of the American Statistical Association*, 65, 1350-1361.