Estimation of the Distribution of Hourly Pay from Household Survey Data: The Use of Missing Data Methods to Handle Measurement Error

Gabriele Beissel-Durrant, Chris Skinner

Abstract

This paper considers the application of missing data methods to a measurement error problem arising in the estimation of the distribution of hourly pay in the United Kingdom, using data from the Labour Force Survey. Errors in the measurement of hourly pay 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..

S[3]RI Methodology Working Paper M04/08

Estimation of the Distribution of Hourly Pay from Household Survey Data: The Use of Missing Data Methods to Handle Measurement Error

Gabriele Beissel-Durrant and Chris Skinner

Summary. This paper considers the application of missing data methods to a measurement error problem arising in the estimation of the distribution of hourly pay in the United Kingdom, using data from the Labour Force Survey. Errors in the measurement of hourly pay 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; nearest neighbour imputation; predictive mean matching; propensity score weighting.

1. Introduction

A national minimum wage was introduced in the United Kingdom (UK) in April 1999 and there is considerable interest in how the lower end of the distribution of hourly pay has changed since then, both overall and within subgroups, such as by gender. The UK Labour Force Survey (LFS) provides an important source of estimates of this distribution (Stuttard and Jenkins 2001). A major problem with the use of household surveys to produce such estimates is the difficulty in measuring hourly pay accurately (Rodgers, Brown and Duncan 1993; Moore, Stinson and Welniak 2000). Measurement error may lead to biased estimates of distribution functions, especially at the extremes (Fuller 1995). For example, the bold line in Figure 1 represents a standard estimate of the lower end of the distribution function of hourly pay using LFS data from the June-August 1999 ignoring measurement error. We suggest that this estimate is seriously biased upwards and that improved estimates, using methods to be described in this paper, are given by the three lower lines. These 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.

[Figure 1 about here]

When a variable is measured with error, it is sometimes possible, as in our application, to measure the variable more accurately for a subsample. In these circumstances, if we assume that the variable measured accurately on the subsample is the true variable, inference about the distribution of this variable becomes a missing data problem. The variable measured erroneously on the whole sample is treated as an auxiliary variable. The case when the subsample is selected using a randomised scheme is well studied and referred to as double sampling or two phase sampling (e.g. Tenenbein 1970). In this case, unbiased estimates can be constructed from the subsample alone, but use of data on the correlated proxy variable for the whole sample may improve efficiency. See, for example, Luo, Stokes and Sager (1998). In the application in this paper, the selection of the subsample is not randomised and 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. Because the aim is to estimate a distribution function, which is unlikely to follow exactly a standard parametric form, we avoid approaches which make parametric assumptions about the true distribution, as for example in Buonaccorsi (1990). It is also desirable in our application to avoid strong assumptions about the measurement error model, for example that it is additive with zero mean and constant variance as in the SIMEX method of Luo et al. (1998). Instead, the novel feature of this paper is to consider the application of various imputation and weighting methods from the missing data literature to this measurement error problem. The aim of the paper is to investigate how to design these methods to improve point estimation of the distribution function of hourly pay, in terms of bias, efficiency and robustness to model assumptions.

The basic measurement error problem considered in this paper was described by Skinner, Stuttard, Beissel-Durrant and Jenkins (2002), who also proposed the use of one imputation method. 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 Office for National Statistics in the United Kingdom as a new approach to producing low pay estimates.

The paper is structured as follows. The application and the estimation problem are introduced in section 2. Imputation and weighting 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 6. Section 7 discusses the application of the missing data methods to the Labour Force Survey. Some concluding remarks are given in section 8.

2. The Estimation Problem

Our aim is to estimate the distribution of hourly pay from LFS data. This inference problem requires consideration of both (i) sampling and unit nonresponse of employees and (ii) measurement error and item nonresponse for hourly pay. We outline the basic set-up for both (i) and (ii) in this section. The main focus of the paper will be the choice of methods to address (ii). Standard procedures will be used to handle (i).

The LFS 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 but not otherwise by geography. Each selected household is retained in the sample for interview on five successive quarters and then rotated out and replaced. The questions underlying the hourly pay variables, described below, are asked in just the first and fifth interviews, giving information on hourly pay on about 17,000 employees per quarter. Survey weights are constructed to compensate for differential unit nonresponse (ONS, 1999).

The traditional method of measuring hourly pay in the LFS is (a) to ask employees questions about their main job to determine earnings over a reference period, (b) to ask questions to determine hours worked over the reference period and (c) to divide the result of (a) by the result of (b). We refer to the result of (c) as the *derived hourly pay* variable. This is the variable used to produce the bold line in Figure 1. Skinner et al. (2002) describe and provide empirical evidence of many sources of measurement error in this variable. A more recent method of measuring hourly pay is first to ask whether the respondent is paid a fixed hourly rate and then, if the answer is positive, to ask respondents what this (basic) rate is. We refer to the resulting measure of hourly pay as the *direct variable.* Skinner et al. (2002) conclude from their study that the direct variable measures hourly pay much more accurately than the derived variable and a key working assumption of this paper is that the direct variable measures hourly pay without error.

The problem with the direct variable 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 may be expected to be 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 first jobs.

This paper addresses the following missing data problem. We wish to estimate the distribution of hourly pay defined as:

$$F(y) = N^{-1}\sum_{i\in U} I(y_i < y) \qquad (1)$$

where U is the population of N (first) jobs, y_i is (basic) hourly pay for job i and y may take any specified value. 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 $r_i = 1$, where s is the set of (first) jobs for unit respondents in the sample drawn from U, y_i^* is the value of the derived variable, y_i is the value of the direct variable assumed identical to the hourly pay variable of interest, $r_i = 1$ if y_i is measured and $r_i = 0$ if not and x_i is a vector of other variables measured in the survey.

We assume that inference from the sample s to the population U can be made using standard methods of survey sampling. Our primary concern is with the missingness of y_i. We consider two approaches to handle this missingness:

(i) imputation of y_i for cases where $r_i = 0$ ($i \in s$), using the values y_i^* and x_i as auxiliary information;

(ii) weighting of an estimator based upon the 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.

3. Imputation Approaches

We shall construct imputation methods based upon the assumption that the population values (y_i , y_i^*, x_i, r_i), $i \in U$, are independently and identically (IID) distributed. To allow for the LFS sampling design and unit nonresponse, we propose to incorporate the survey weights in the resulting point estimator of $F(y)$, in the same way that a pseudo-likelihood approach (Skinner, 1989) weights estimators based upon an IID assumption. We do not attempt to take account of the weights or the complex design directly in the imputation methods.

Under the IID assumption and the assumption that sampling is ignorable (that is that the distribution of (y_i , y_i^*, x_i, r_i) is the same whether or not $i \in s$), if it were possible to observe y_i for $i \in s$,

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

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 $s = \{1,...,n\}$. We assume that this estimator remains unbiased under the actual sampling design and unit nonresponse if the mean in (2) is weighted by the survey weights. The IID assumption used in the remainder of this section may be interpreted as holding conditional on inclusion in s , with the implicit assumption that survey weighting will also be required to handle the selection of *s* from *U*.

To address the problem that y_i is missing when $r_i = 0$, we first consider a single imputation approach where y_i is replaced in (2) by a single 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. We assume that y_i^I is determined in a specified way using the data $D = \{[y_i^*, x_i, r_i;\ i \in s], [y_i;\ r_i = 1, i \in s]\}$ and perhaps a stochastic mechanism. The resulting estimator of $F(y)$ is

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

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 $[y_i^I \mid r_i = 0]$, is the same as the conditional distribution $[y_i \mid r_i = 0]$. However, since y_i is only observed when $r_i = 1$, the data provide no direct information about $[y_i \mid r_i = 0]$ without further assumptions. We consider two possible assumptions. The first is common in the missing data literature (Little and Rubin 2002).

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

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 ($r_i = 0$), which may be expressed as follows.

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 and is the one which we shall make. We shall use the second assumption in the simulation study in section 6 to assess robustness of MAR-based procedures. Inference under the second assumption could be made under strong assumptions on the measurement error model, for example the additive error assumption in methods in Carroll, Ruppert and Stefanski (1995, sect. 12.1.2.) and Luo et al. (1998). It does not appear straightforward to make inference under the second assumption for a measurement error model which is realistic for our application and we do not pursue this possibility further in this paper. The plausibility of these assumptions is discussed further in Skinner et al. (2002).

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

$$[\, y_i^I \mid y_i^*, x_i, r_i = 0] = [\, y_i \mid y_i^*, x_i, r_i = 1]. \tag{4}$$

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. We consider representing the conditional distribution $[\, y_i \mid y_i^*, x_i, r_i = 1]$ by a parametric regression model:

$$g(\, y_i) = h(\, y_i^*, x_i; ß) + e_i ,\; E(e_i \mid y_i^*, x_i\,) = 0 \tag{5}$$

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

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

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 may be expected to reduce the variation in $F(Y)$ artificially (Little and Rubin 2002, p. 64). This effect might be avoided by taking $y_i^I = g^{-1}[h(\, y_i^*, x_i; \hat{ß}) + \hat{e}_i]$, where $\hat{e}_i$ is a randomly selected empirical residual (Little and Rubin 2002, p. 65). Our experience is, however, that this approach fails to generate imputed values which reproduce the 'spiky' behaviour of hourly pay distributions, for example around a minimum wage, and this may lead to bias around these spikes. We prefer therefore to consider 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, as reported by the donor respondent, and will thus respect the spiky behaviour these values display. The basic donor imputation method we consider is predictive mean matching (Little 1988), that is nearest neighbour imputation with respect to $\hat{y}_i$, i.e.

$$|\, \hat{y}_i - \hat{y}_{d(i)} | = \min_{j: r_j = 1} |\, \hat{y}_i - \hat{y}_j \,| \tag{7}$$

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 certain conditions hold. The four conditions are that: y_i is missing at random (MAR) conditional on $z_i = g^{-1}[h(y_i^*, x_i; ß)]$, where $ß = \text{plim}(\hat{ß})$; the conditional expectation of y_i given z_i is monotonic and continuous in z_i; z_i and $E(y_i \mid z_i)$ have finite third moments; and the probability of response given z is bounded above zero. These conditions seem plausible provide 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 good proxy for y_i and if we restrict attention to the lower part of the pay distribution. In addition, Chen and Shao's (2000) result needs to be adapted for the fact that the nearest neighbour is defined with respect to $\hat{ß}$ whereas the above conditions are with respect to $ß$. This approximation also seems plausible since close neighbours with respect to $\hat{y}_i = g^{-1}[h(y_i^*, x_i; \hat{ß})]$ should also be close neighbours with respect to $z_i = g^{-1}[h(y_i^*, x_i; ß)]$. 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 smooth 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 of $\hat{y}_i$ 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

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

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 $y_i^{I(m)}$, $m = 1,...,M$, determined 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 $\tilde{F}^{(m)}(y)$, or equivalently is obtained by multiplying the weight for each imputed value by a factor $1/M$. 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 and to achieve consistent variance estimation would raise further issues such as the effect of cluster sampling of adults within households. 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 $\sum_{s_1} w_i = n$. The weight w_i may be multiplied by the survey weight to allow for unit nonresponse. Other choices of weight w_i may also be considered. In particular, we may 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 in this paper by Dickens and Manning (2002). This 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 $[r_i | y_i^*, x_i]$ and some regularity conditions, such as those described in section 3 for the imputed estimator. Note that the need to model $[r_i | y_i^*, x_i]$ replaces the need to model $[y_i | y_i^*, x_i]$ 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 $q = E(u_i)$. We consider the imputation approach first and suppose that y_i depends upon y_i^* and x_i only via $z_i = g^{-1}[h(y_i^*, x_i; b]$ and that y_i is missing at random given z_i. Ignoring the difference between b and $\hat{b}$ for large samples we consider nearest neighbour imputation with respect to z_i. As in (9) the imputed estimator of q may be expressed as

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

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

$$Pr(r_i = 1 | y_i^*, x_i) = p(z_{PSi}). \qquad (11)$$

Just as we ignored the difference between b and $\hat{b}$, we ignore error in estimating $p(z_{PSi})$ and write $w_{PSi} = p(z_{PSi})^{-1}$.

The imputation and propensity score weighting approaches may be expected to give similar estimation results if z_i and z_{PSi} 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, then 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 other nearest neighbour imputation approaches considered in section 3, the weight $w_i = 1 + d_i / M$ may be interpreted as a local (with repect to z_i) nonparametric estimate of $\Pr(r_i = 1 | z_i)^{-1}$ and thus may be expected to lead to similar estimation results to propensity score weighting if z_i and z_{PSi} 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_{PSi} 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 upon propensity score weighting, as also discussed by David et al. (1983) and Rubin (1996, sect. 4.6). To investigate this further, let us now make the MAR assumption and the other assumptions in sections 3 and 4 upon which the approaches are based. In this case 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), under their regularity conditions, that the variance of $\hat{q}_{IMP}$ may be approximated for large n by

$$\text{var}(\hat{q}_{IMP}) \approx n^{-2} E[\sum_{s_1} w_i^2 V(u_i | z_i)] + n^{-1} V[y(z_i)] \quad (12)$$

where $y(z_i) = E(u_i | z_i)$. 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 using

$$V[y(z_i)] = s^2 - E[V(u_i | z_i)], \quad (13)$$

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

$$E[n^{-1} \sum_{s_1} w_i V(u_i | z_i)] = E[V(u_i | z_i)] + o_p(n^{-1/2}). \quad (14)$$

It follows that to the same order of approximation as in (12)

$$\text{var}(\hat{q}_{IMP}) \approx n^{-1} s^2 + n^{-2} E[\sum_{s_1} (w_i^2 - w_i) V(u_i | z_i)]. \quad (15)$$

Note that $w_i^2 - w_i = (d_i / M)(1 + d_i / M) \geq 0$. This expression may be interpreted from both 'missing data' and 'measurement error' perspectives. From a missing data perspective, the first term in (15) is just the variance of $\hat{q}$ 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^* becomes a better measure of y_i and $V(u_i \mid z_i)$ approaches zero. In this case, the second term also approaches zero and $\hat{q}_{IMP}$ becomes 'fully efficient', i.e. its variance approaches s^2 / n.

Let us now consider propensity score weighting. We make the corresponding assumption that y_i is missing at random given z_{PSi}. Linearising the ratio in (9) and using the fact that $E(\sum_{s_1} w_{PSi}) = n$ we may write

$$\mathrm{var}(\hat{q}_{PS}) \approx n^{-2} \mathrm{var}[\sum_{s_1} w_{PSi}(u_i - q)] \qquad (16)$$

$$= n^{-1} E[w_{PSi}(u_i - q)^2]$$

which may be expressed alternatively as

$$\mathrm{var}(\hat{q}_{PS}) \approx n^{-2} E[\sum_{s_1} w_{PSi}^2 V(u_i \mid z_{PSi})] + n^{-1} E\{w_{PSi}[y(z_{PSi}) - q]^2\}. \qquad (17)$$

To compare the efficiency of weighting and imputation it is convenient to use (13) and (14) (which hold also with w_{PSi} in place of w_i) to obtain

$$\mathrm{var}(\hat{q}_{PS}) \approx n^{-1} s^2 + n^{-2} E[\sum_{s_1} (w_{PSi}^2 - w_{PSi}) V(u_i \mid z_{PSi})]$$

$$+ n^{-1} E\{\sum_{s_1} [w_{PSi} - 1][y(z_{PSi}) - q]^2\}. \qquad (18)$$

Note that, in comparison with (15), this involves a third term, which does not necessarily converge to zero as y_i^* approaches y_i and $V(u_i \mid z_{PSi}) \to 0$. 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 (15) when $V(u_i \mid z_i)$ and $V(u_i \mid z_{PSi})$ are constant and equal, since, recalling that $\sum_{s_1} w_i = E(\sum_{s_1} w_{PS\, i}) = n$, these second terms are primarily determined by the variances of the weights w_i and w_{PSi}, and, provided M is sufficiently large, we may expect w_i to display less variation than w_{PSi}, as argued above. In general, however, it does not appear that $\hat{q}_{IMP}$ is necessarily more efficient than $\hat{q}_{PS}$ and we look to the simulation study in section 6 for numerical evidence.

Let us finally consider the impact of departures from the MAR assumption. Under small measurement error asymptotics where $V(u_i \mid z_i) \rightarrow 0$ and $y_i^I \rightarrow y_i$, the imputation approach will provide consistent inference about q 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. Simulation Study

The aim of the study is to generate independent repeated samples $s^{(h)}$, $h = 1,\ldots,H$, with realistic values y_i, y_i^*, x_i, r_i, $i \in s^{(h)}$, 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 jobs 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 effective assumption that the population size is infinite seems reasonably given the small sampling fraction of the LFS. The assumption of (simple) random sampling neglects the clustering of the sample into households, although the impact of this simplification on the relative properties of estimators is expected to be slight. 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 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 and H=1000 and generated the values of y_i, y_i^* and r_i for each sample $s^{(h)}$ 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 7000 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 $\hat{q}_t^{(h)}$ of two parameters $(t = 1,2)$ were obtained for each sample $s^{(h)}$,

$q_1 =$ proportion with pay below the national minimum wage (=£3.00 per hour age 18-21, £3.60 per hour aged 22+)

$q_2 =$ proportion with pay between minimum wage and £5/hour.

The true values are q_1 =0.056 and q_2 = 0.185. The bias and standard error were estimated as

$\widehat{bias}(\hat{q}_t) = \bar{q}_t - q_t$ and $\widehat{s.e.}(\hat{q}_t) = [H^{-1}\sum_{h=1}^{H}(\hat{q}_t^{(h)} - \bar{q}_t)^2]^{1/2}$, where $\bar{q}_t = H^{-1}\sum_h q_t^{(h)}$.

We first compare results for the alternative imputation approaches. Table 1 presents estimates of the biases of estimators of q_1 and q_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, p. 227) for stochastic methods of imputation, such as the HDI methods, compared to deterministic methods, such as nearest neighbour imputation, when estimating distributional parameters.

[Table 1 about here]

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 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 2 about here]

We next compare the NN10 imputation approach with propensity score weighting. 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 q_1 or q_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 noticeably greater for the weighting approach. Corresponding estimated standard errors of $\hat{q}_1$ and $\hat{q}_2$ are given in Table 4. These also tend to be greater for the weighting approach with the increase of mean squared error 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.

[Table 3 and 4 about here]

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 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 $\hat{q}_1$ is as expected from arguments in Dickens and Manning (2002) and Skinner et al. (2002). The relative bias of 5% of the NN10 approach does not, however, appear to make the resulting estimates unusable.

[Table 5 about here]

7. Application to the Labour Force Survey

In this section nearest neighbour imputation, hot deck imputation within classes and propensity score weighting are applied to LFS data. 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 differences between the missing data methods are much smaller. Note that the June-August 1999 quarter is subject to a lower response rate. It was found that for consecutive quarters, which are subject to about 43% response rate, weighting and imputation lead to very similar estimates of low pay proportions. 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. The estimates in both Figure 1 and Table 6 employ survey weights. 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.

[Table 6 about here]

8. 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.

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 3.

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.

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

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

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. (2002) Has the National Minimum Wage reduced UK wage inequality?, Discussion Paper 533, Centre for Economic Performance, London School of Economics, http://cep.lse.ac.uk/pubs/download/dp0533.pdf .

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

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.

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

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, 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.

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.

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 C. J. Skinner, D. Holt and T. M. F. Smith eds., *Analysis of Complex Surveys*, 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.

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

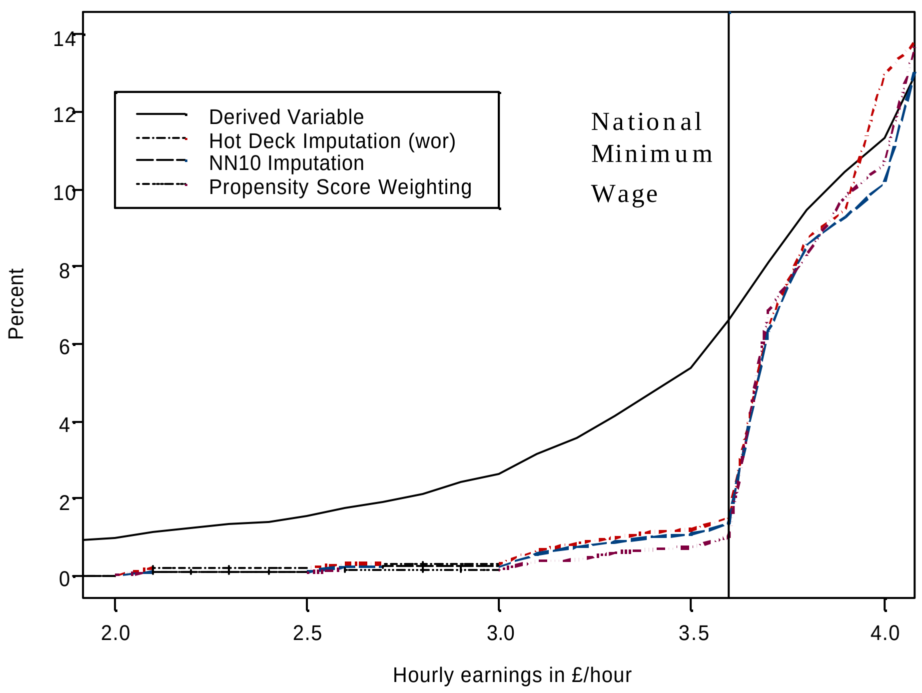

Table 1. Simulation Estimates of Biases of Estimators of q_1 and q_2 for Different Imputation Methods, Assuming MAR and Correct Covariates.

Imputation Method	Bias of $\hat{q}_1$	Rel. Bias of $\hat{q}_1$	Bias of $\hat{q}_2$	Rel. Bias of $\hat{q}_2$
NN1	$1.2*10^{-4}$ ($0.9*10^{-4}$)	0.2 %	$0.9*10^{-4}$ ($1.7*10^{-4}$)	0.0 %
NN1P[1]	$4.4*10^{-4}$ ($2.6*10^{-4}$)	0.8 %	$0.3*10^{-4}$ ($5.1*10^{-4}$)	0.0 %
NN2	$0.6*10^{-4}$ ($8.5*10^{-4}$)	0.1 %	$1.6*10^{-4}$ ($1.5*10^{-4}$)	0.0 %
NN2(4)	$1.4*10^{-4}$ ($0.9*10^{-4}$)	0.2 %	$-2.5*10^{-4}$ ($1.5*10^{-4}$)	-0.1 %
NN10	$0.2*10^{-4}$ ($6.5*10^{-4}$)	0.0 %	$-1.2*10^{-4}$ ($1.5*10^{-4}$)	-0.1 %
NN10(20)	$0.2*10^{-4}$ ($0.8*10^{-4}$)	0.0 %	$0.7*10^{-4}$ ($1.5*10^{-4}$)	0.0 %
HDIWR10	$2.8*10^{-4}$ ($0.7*10^{-4}$)	0.5 %	$26.2*10^{-4}$ ($1.5*10^{-4}$)	1.4 %
HDIWOR10	$2.5*10^{-4}$ ($0.7*10^{-4}$)	0.4 %	$28.0*10^{-4}$ ($1.2*10^{-4}$)	1.5 %
ABB10	$4.6*10^{-4}$ ($0.8*10^{-4}$)	0.8 %	$29.8*10^{-4}$ ($1.5*10^{-4}$)	1.6 %

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 q_1 and q_2 for Different Imputation Methods, Assuming MAR and Correct Covariates.

Imputation Method	$s.e.(\hat{q}_1)$	$s.e.(\hat{q}_2)$	$\frac{V(\hat{q}_1)}{V_{NN1}(\hat{q}_1)}$	$\frac{V(\hat{q}_2)}{V_{NN1}(\hat{q}_2)}$
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.

Table 3. Simulation Estimates of Biases of Estimators of q_1 and q_2 for Nearest Neighbour Imputation (NN10) and Propensity Score Weighting, Assuming MAR and Correct and Misspecified Covariates.

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

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 q_1 and q_2 for Nearest Neighbour Imputation (NN10) and Propensity Score Weighting, Assuming MAR and Correct and Misspecified Covariates.

Method	**Assumed Covariates**	$s.e.(\hat{q}_1)$	$s.e.(\hat{q}_2)$	**MSE($\hat{q}_1$)**	**MSE($\hat{q}_2$)**
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 q_1 and q_2 for Nearest Neighbour Imputation (NN10) and Propensity Score Weighting. Under the (non-MAR) Common Measurement Error Model.

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

Table 6. Estimates of q_1 and q_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) $\hat{q}_1$ (%)	(Weighted) $\hat{q}_2$ (%)
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