IMPUTATION VS. ESTIMATION OF FINITE POPULATION DISTRIBUTIONS

R. L. CHAMBERS

ABSTRACT

Estimates of the distribution of hourly wage rates for employees are an important output for a national statistics agency. However, many employees are not paid by the hour and so their hourly wage rate data are effectively missing in a survey that attempts to collect this information. A standard approach in this situation is to impute these missing values using derived measures of this wage rate based on salary and hours worked data also collected in the survey. This paper contrasts this imputation approach with direct estimation of the wage rate distribution using the derived wage rate variable as an auxiliary. In particular, we focus on data obtained in the 2002 UK New Earnings Survey and use simulation based on actual and derived hourly wage rate data collected in this survey to compare two imputation approaches, one based on substituting the derived wage rate values for the missing actual values, the other using nearest neighbour imputation based on the derived wage rate, with two estimation approaches that use this variable as an auxiliary. The first of these is a semi-parametric extension of the Chambers and Dunstan (1986) estimator of the finite population distribution function, the other is a calibrated spline-based estimator of this function recently suggested by Harms and Duchesne (2004). Our conclusion is that an approach based on the semi-parametric estimator is best for these data. However, confidence interval estimation remains an open problem.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M05/06

Imputation vs. Estimation of Finite Population Distributions

R. L. Chambers

Southampton Statistical Sciences Research Institute
University of Southampton
Highfield, Southampton SO17 1BJ
United Kingdom

Abstract

Estimates of the distribution of hourly wage rates for employees are an important output for a national statistics agency. However, many employees are not paid by the hour and so their hourly wage rate data are effectively missing in a survey that attempts to collect this information. A standard approach in this situation is to impute these missing values using derived measures of this wage rate based on salary and hours worked data also collected in the survey. This paper contrasts this imputation approach with direct estimation of the wage rate distribution using the derived wage rate variable as an auxiliary. In particular, we focus on data obtained in the 2002 UK New Earnings Survey and use simulation based on actual and derived hourly wage rate data collected in this survey to compare two imputation approaches, one based on substituting the derived wage rate values for the missing actual values, the other using nearest neighbour imputation based on the derived wage rate, with two estimation approaches that use this variable as an auxiliary. The first of these is a semi-parametric extension of the Chambers and Dunstan (1986) estimator of the finite population distribution function, the other is a calibrated spline-based estimator of this function recently suggested by Harms and Duchesne (2004). Our conclusion is that an approach based on the semi-parametric estimator is best for these data. However, confidence interval estimation remains an open problem.

Key Words Missing data; Nearest neighbour imputation; Calibrated estimation; Wage distribution estimation.

1. Introduction

Much of the literature on survey sampling inference focuses on finite population totals and means. However, users of survey data are often more interested in the finite population distribution of a survey variable, and measures (e.g. medians, quartiles, percentiles) that characterise the shape of this distribution. In this paper we consider a particular application where the primary target of inference is a finite population distribution function, but where there are two radically different ways one can tackle the problem of making an inference about this function. Both use available auxiliary information, but in very different ways. The first uses this information to impute the population values defining the function, while the second uses this information to estimate the actual value of this function. In what follows we contrast these two approaches by means of a simulation study based on a real life data set. Our conclusion is that for these data it is preferable to estimate provided the sampling method is uninformative. If this is not the case, then imputation is preferable. However, in both cases the method of estimation and the method of imputation need to be carefully specified.

To start, we provide the motivation for our study. Employment law in the UK dictates that all employees above a certain age must be paid an hourly wage greater than or equal to a set minimum value. This "minimum wage" itself is subject to change over time, reflecting the impact of inflationary pressures on individual purchasing power. One important consideration in the government's assessment of whether to make any changes to the minimum wage is the impact of such a change on the national wage bill. This in turn requires knowledge of the distribution of wage rates across the economy, particularly the lower tail of this distribution.

Till recently, an important source of information about this distribution was the New Earnings Survey (NES) carried out by the UK Office for National Statistics. This was a large-scale annual survey of employees in the UK business sector that collected data on salaries, hours worked and hourly rates of pay. From 2004 the NES was replaced by the Annual Survey of Hours and Earnings (ASHE), which has essentially the same remit. We confine our analysis in this paper, however, to data collected in the 2002 round of NES, since data collected in ASHE are expected to be similar.

A key objective of NES was measurement of the distribution of Y = hourly pay rates for all UK employees. However, these hourly rates cannot be obtained from all responding employees since many are not paid by the hour. In such cases, it is possible to calculate an implicit hourly rate (X = derived rate) based on total earnings and hours worked, which are available for all responding employees. Unfortunately, Y and X are not the same, even when both are available. This is clearly illustrated in Table 1, which shows the distributions of these variables for NES respondents providing values for both, as well as for respondents that provide values for X alone. In particular, we see that values of X in the latter group tend to be considerably larger than values of X in the former group. In Table 2 we focus on the marginal distributions of Y and X for a subset of the NES respondents that provided data for both. This subset is defined by excluding all respondents that provided these data but had implausibly small values for either Y or X, or where either of these values was very large. Figure 1 is the scatterplot of Y and X values underpinning the data contributing to Table 2. Here we see that although there are clearly many employees where Y and X are very similar, there is also a large amount of variability.

Tables 1 and 2 and Figure 1 about here.

2. Imputing the NES Data

Since it is impossible to distinguish between an employee who has an hourly wage but does not provide a value for Y (i.e. a non-respondent) and one who cannot because he or she is not paid an hourly wage, we assume from now on that all missing values of Y are due to inability to provide this value. This is reasonable, since NES sample s is essentially a simple random sample of national insurance contributor numbers, and the data are actually provided by the employer. Let s_1 denote the n_1 sampled employees that provide data for Y and X and let s_2 denote the n_2 sampled employees that provide data for X alone. If data on Y were available for all $n = n_1 + n_2$ sampled employees, the estimator of the proportion of wage earners with hourly pay rates less than or equal to t would be

$$\hat{F}_s(t) = n^{-1}\left[\sum_{s_1} I(y_i \le t) + \sum_{s_2} I(y_j \le t)\right] \tag{1}$$

where $I(u)$ takes the value 1 if u is correct and is zero otherwise.

Since the second term on the right is unknown, (1) cannot be calculated. We therefore consider three options that one might adopt at this stage:

1. We could do nothing, i.e. we do not impute and hence ignore the information in s_2. In this case we estimate using the information in s_1, replacing (1) by the available data estimator

$$\hat{F}_{s1}(t) = n_1^{-1}\sum_{s_1} I(y_i \le t). \tag{2}$$

2. We could impute each missing value of Y by the corresponding derived value X, leading to the substitution estimator

$$\hat{F}_{sub}(t) = n^{-1}\left[\sum_{s_1} I(y_i \le t) + \sum_{s_2} I(x_j \le t)\right]. \tag{3}$$

   This makes sense if one defines Y as X when no hourly wage is being paid.

3. We could impute each missing value of Y by making a random draw from an estimate of the conditional distribution of $Y|X$, leading to the imputation estimator

$$\hat{F}_{imp}(t) = n^{-1}\left[\sum_{s_1} I(y_i \le t) + \sum_{s_2} I(Y^*(x_j) \le t)\right] \tag{4}$$

   Here $Y^*(x_j)$ denotes the random draw for the s_2 unit with $X = x_j$.

There are a variety of ways of implementing imputation option 3 above. In this paper we adopt a method based on nearest neighbour imputation. That is, we find the s_1 unit with X value $x_{near}(x_j)$ that is closest to x_j. Let $y_{near}(x_j)$ denote the value of Y associated with this unit. We then choose $Y^*(x_j)$ by making a random draw from the convolution of the empirical s_1 distribution of Y with a "smearing" distribution centred at $y_{near}(x_j)$.

It is interesting to observe that the substitution estimator can be viewed as a close approximation to a nearest neighbour imputation-based estimator. However, in this case we find the s_1 unit with Y (rather than X) value $y_{near}(x_j)$ closest to x_j. A "random draw" version of the substitution estimator is easily defined using the same type of smearing distribution procedure as described above.

It is clear that one can make multiple independent draws (with replacement) from the convolution distribution that defines (4). This suggests that we average the resulting single imputation-based estimates over these draws in order to reduce the Monte Carlo variability associated with the single draw estimator (4). It is not difficult to see that the limiting form of this averaged estimator is

$$\hat{F}_{imp}^{\infty}(t) = n^{-1}\sum\nolimits_{s_1}\left(1+\sum\nolimits_{j\in s_2} p_{ij}\right)I(y_i \le t). \tag{5}$$

Here $\{p_{ij}; i \in s_1\}$ are the (known) probabilities defining the convolution distribution used to generate the imputed value $Y^*(x_j)$ in s_2.

3. Estimation as an Alternative to Imputation

An alternative to estimation using imputed values is to estimate the "complete response" distribution function (1) using the sample X values as auxiliary information. Again, there are a variety of ways this can be done. However, we focus on two quite distinct approaches that are representative of the model-based and design-based approaches to this estimation problem. The model-based approach is based on the predictor of the finite population distribution function suggested by Chambers and Dunstan (1986), while the design-based approach is based on the calibrated estimator of this function suggested by Harms and Duchesne (2004).

In what follows we motivate each of these approaches in turn. We do this in the context of "standard" survey sampling. That is, given a population U of size N from which a sample s of size n has been taken, we focus on estimation of the finite population distribution function

$$F_N(t) = N^{-1}\sum\nolimits_U I(y_i \le t) = N^{-1}\left(\sum\nolimits_s I(y_i \le t) + \sum\nolimits_{U-s} I(y_j \le t)\right). \tag{6}$$

3.1 Model-Based Estimation of a Finite Population Distribution

Chambers and Dunstan (1986, hereafter referred to as CD) assume that values of an auxiliary variable x are available for all units in the population, the first two moments of $y_i \mid x_i$ exist for all $i \in U$, distinct population units are independent and sampling is uninformative given the population values $\mathbf{x}_U$ of x. This allows them to write $y_i = \mu(x_i) + \sigma(x_i)\varepsilon_i$ where $\mu(x_i) = E(y_i \mid x_i)$, $\sigma(x_i)$ is a strictly positive function and the ε_i are independent and identically distributed with zero mean and unit variance. The minimum mean squared error predictor of (6) can then be written

$$\begin{aligned}\tilde{F}_N(t) &= N^{-1}\left(\sum\nolimits_s I(y_i \le t) + \sum\nolimits_{U-s} E\left(I(y_j \le t) \mid \mathbf{x}_U\right)\right)\\ &= N^{-1}\left(\sum\nolimits_s I(y_i \le t) + \sum\nolimits_{U-s} \Pr\left(\mu(x_j) + \sigma(x_j)\varepsilon_j \le t\right)\right).\end{aligned}$$

CD used this representation to motivate the estimator

$$\hat{F}_{CD}(t) = N^{-1}\left(\sum_s I(y_i \le t) + \sum_{U-s} \widehat{\Pr}\left(y_j \le t \mid \mathbf{x}_U\right)\right) \tag{7}$$

where

$$\widehat{\Pr}\left(y_j \le t \mid \mathbf{x}_U\right) = n^{-1}\sum_{i\in s} I\left(\hat{\mu}(x_j) + \hat{\sigma}(x_j)\left(\frac{y_i - \hat{\mu}(x_i)}{\hat{\sigma}(x_i)}\right) \le t\right). \tag{8}$$

Here $\hat{\mu}(x)$ and $\hat{\sigma}(x)$ denote sample-based estimates of $\mu(x)$ and $\sigma(x)$. The CD estimator (7) is highly efficient provided $\hat{\mu}(x)$ and $\hat{\sigma}(x)$ are "good" estimators of $\mu(x)$ and $\sigma(x)$ over the range of non-sample *x*-values. However it can be biased if either is misspecified.

There are two basic approaches one can take to implementing the CD estimator in situations where model specification uncertainty exists. The first is to use nonparametric estimates of $\mu(x)$ and $\sigma(x)$ in (8). See Dorfman and Hall (1993) and Lombardia *et al.* (2004). The second, and conceptually simpler, approach is just to replace (8) by a "local", rather than "global" mean, leading to the semi-parametric estimator

$$\hat{F}_{CDL}(t) = N^{-1}\left(\sum_s I(y_i \le t) + \sum_{U-s}\left\{\frac{\sum_{i\in s} w_i(x_j) I\left(y_i - \hat{\mu}(x_i) \le t - \hat{\mu}(x_j)\right)}{\sum_{i\in s} w_i(x_j)}\right\}\right) \tag{9}$$

where $\hat{\mu}(x)$ is a reasonable parametric estimate of $\mu(x)$ and the weights $w_i(x_j)$ are local weights, i.e. they satisfy $\|x_i - x_j\| \le \|x_k - x_j\| \Rightarrow w_i(x_j) \ge w_k(x_j)$. We shall use (9). There are many ways the local weights in this estimator can be defined. We use the simple specification

$$w_i(x_j) = I\left(\|x_i - x_j\| \le \frac{range(x)}{f}\right)$$

provided $\sum_{i\in s} w_i(x_j) > 5$, otherwise we set $w_i(x_j) = 1$. Note that $1/f$ plays the role of a bandwidth here and the CD estimator corresponds to $f = 1$. A straightforward way of choosing *f* is then via an ordered half-sample cross-validation procedure, defined as follows:

- Order the sample *x*-values: $x_{(1)}, x_{(2)}, x_{(3)}, \cdots, x_{(n-1)}, x_{(n)}$;
- Create two sets $E = \{x_{(1)}, x_{(3)}, \cdots\}$ and $V = \{x_{(2)}, x_{(4)}, \cdots\}$;
- For given *f* and *t* compute (9) treating *E* as the "sample" and *V* as the "non-sample". Denote this estimate by $\hat{F}^{(n)}_{CDL}(t)$;
- Choose the value of *f* that minimises the integrated squared distance $\sum_{g=1}^{G}\left\{\hat{F}^{(n)}_{CDL}(t_g) - n^{-1}\sum_s I(y_i \le t_g)\right\}^2$ over a pre-specified grid $\{t_g\}$ of *t*-values.

In some circumstances the fit at some *t*-values is more important than at others, and we use a weighted integrated squared distance criterion

$$\sum_{g=1}^{G} \theta_g \left\{ \hat{F}_{CDL}^{(n)}(t_g) - n^{-1} \sum_s I(y_i \le t_g) \right\}^2$$

where the $\{\theta_g; g = 1,\ldots,G\}$ are pre-specified weights reflecting the relative importance of each *t* value on the grid.

Without loss of generality let $j,k = 1,\ldots,N-n$ index the non-sampled units in the population when they are ordered by increasing value of $\hat{\mu}(x)$. In the Annex we describe how a large sample estimator of the prediction variance of (9) can be derived. This is

$$\hat{V} = N^{-2}\left(\hat{V}_s + \hat{V}_{U-s}\right) \tag{10}$$

where

$$\hat{V}_s = \sum_{i \in s} \left[\begin{array}{l} \left\{ \sum_{j=1}^{N-n} w_{ij}^* \left(1 - \hat{P}_{si}(t - \hat{\mu}(x_j))\right) \right\} \left\{ \sum_{j=1}^{N-n} w_{ij}^* \hat{P}_{si}(t - \hat{\mu}(x_j)) \right\} \\ -\sum_{j=1}^{N-n} \sum_{k=1}^{j} w_{ij}^* w_{ik}^* \left\{ \hat{P}_{si}(t - \hat{\mu}(x_k)) - \hat{P}_{si}(t - \hat{\mu}(x_j)) \right\} \end{array} \right]$$

$$\hat{V}_{U-s} = \sum_{U-s} \hat{P}(t, x_j)\left(1 - \hat{P}(t, x_j)\right)$$

$$\hat{P}_{si}(u) = \sum_{m \in s} w_{mi}^* I\left(y_m - \hat{\mu}(x_m) \le u\right)$$

$$\hat{P}(t, x_j) = \sum_{i \in s} w_{ij}^* I\left(y_i - \hat{\mu}(x_i) \le t - \hat{\mu}(x_j)\right)$$

and

$$w_{ij}^* = \frac{w_i(x_j)}{\sum_{m \in s} w_m(x_j)}.$$

The expression for $\hat{V}_s$ above can be extremely time consuming to calculate, especially if *n* and *N* are large. However, "smoothness" of this variance with respect to variation in *x* implies that we can speed up computation of (10) by replacing individual non-sample *x*-values by grouped data (Dunstan and Chambers, 1989). In particular, let $\{x_g; g = 1,\cdots,G\}$ denote the mid-points of a partition of the non-sample *x*-values into *G* groups, with sizes $\{m_g; g = 1,\cdots,G\}$ and such that $x_g < x_h$ when $g < h$. We can then replace $\hat{V}_s$ by

$$\hat{V}_s^{grp} = \sum_{i \in s} \left[\begin{array}{l} \left\{ \sum_{g=1}^{G} m_g w_{ig}^* \left(1 - \hat{P}_{si}(t - \hat{\mu}(x_g))\right) \right\} \left\{ \sum_{g=1}^{G} m_g w_{ig}^* \hat{P}_{si}(t - \hat{\mu}(x_g)) \right\} \\ -\sum_{g=1}^{G} \sum_{h=1}^{g} m_g m_h w_{ig}^* w_{ih}^* \left\{ \hat{P}_{si}(t - \hat{\mu}(x_h)) - \hat{P}_{si}(t - \hat{\mu}(x_g)) \right\} \end{array} \right]$$

and $\hat{V}_{U-s}$ by

$$\hat{V}_{U-s}^{grp} = \sum_{g=1}^{G} m_g \hat{P}(t, x_g)\left(1 - \hat{P}(t, x_g)\right).$$

The final (grouped) estimator of variance is

$$\hat{V}^{grp} = N^{-2}\left(\hat{V}_s^{grp} + \hat{V}_{U-s}^{grp}\right). \tag{11}$$

For small values of G (11) will be a conservative approximation to (10). This is because the second term on the right hand side of $\hat{V}_s^{grp}$ above is non-negative and decreases to zero as the number of groups, G, decreases. That is, the value of (11) increases as the number of groups decreases. This is consistent with the fact that (11) actually corresponds to a large sample variance estimator for a "grouped" version of (9) and that a decreased number of groups implies increased aggregation of the non-sample X-values and hence an increased loss of efficiency for this grouped estimator. We explore the impact of choice of G in section 4.

3.2 Calibrated Estimation of a Finite Population Distribution

This approach is due to Harms and Duchesne (2004), and estimates $F_N(t)$ by the weighted empirical distribution function

$$\hat{F}_{HDw}(t) = \sum\nolimits_s w_i I(y_i \le t) \Big/ \sum\nolimits_s w_i \tag{12}$$

where the weights w_i are calibrated to the known finite population distribution of X. That is, they "recover" this distribution in the sense that, given $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_p < 1$, they satisfy

$$\sum\nolimits_s w_i I(x_i \le Q_x(\alpha_k)) = N\alpha_k$$

and $\sum_s w_i = N$. Here $Q_x(\alpha_k)$ is the α_k-quantile of the population distribution of X and we implicitly assume that (a) the $Q_x(\alpha_k)$ are distinct, and (b) there is at least one sample X-value "between" each of these distinct values. If this is not true, we "drop" calibration constraints (i.e. values α_k) until this condition is satisfied.

Standard results from calibration theory (Deville and Särndal, 1992; Chambers, 1996) can be used to show that if the calibrated weights w_i are chosen to minimise their chi-square distance from the (equal) weights defining the sample empirical distribution function, then the resulting weighted empirical distribution function is equivalent to the simple regression estimator of $F_N(t)$ under the model

$$I(y_i \le t) = \beta_{0t} + \sum_{k=1}^{p} \beta_{kt} I(x_i \le Q_x(\alpha_k)) + error\ . \tag{13}$$

That is, $\hat{F}_{HDw}(t)$ is a P-spline estimator of $F_N(t)$ with knots defined by the calibration constraints.

Furthermore, given (13) it is not difficult to derive the form of the calibrated weights in (12). Define $z_{ki} = I(x_i \le Q_x(\alpha_k))$ and put $\mathbf{Z}_k = (z_{ki}; i = 1, \cdots, N)$ so $\mathbf{Z} = [\mathbf{1}\, \mathbf{Z}_1 \cdots \mathbf{Z}_p]$ denotes the population matrix of values of these variables (**1** denotes a *N*-vector of ones). The weights w_i are then defined by the vector

$$\mathbf{w}_s = (w_i) = \mathbf{1}_s + \mathbf{Z}_s(\mathbf{Z}_s'\mathbf{Z}_s)^{-1}\mathbf{Z}_{U-s}'\mathbf{1}_{U-s} \tag{14}$$

where subscripts of *s* and *U-s* denote appropriate sample/non-sample partitions of population vector/matrix quantities.

Before the Harms and Duchesne (HD) estimator (12) can be used in practice one needs to decide on the calibration constraints to use in calculating the weights (14). The same ordered half-sample cross validation procedure as described earlier for choice of the bandwidth *f* to use in (9) can be used for this purpose. In particular we use this procedure to determine these calibration constraints by assuming that the corresponding α-values are equally spaced and span the (0,1) interval. The CV procedure can then be used to decide on an optimal number *f* of such equally spaced α-values.

Variance estimation for the HD estimator (12) can be carried out using standard results from model-assisted/model-based sample survey theory. We develop a model-based variance estimator, noting that it is essentially the same as the (model-assisted) variance estimator suggested by Harms and Duchesne (2004).

Since the HD estimator is a linear estimator, its prediction variance is

$$Var\left(\hat{F}_{HDw}(t) - F_N(t)\right) = N^{-2}\left(\sum\nolimits_s (w_i - 1)^2 V_i(t) + \sum\nolimits_{U-s} V_j(t)\right)$$

where $V_i(t) = \Pr(y_i \le t)\left(1 - \Pr(y_i \le t)\right)$. Suppose that the spline model (13) is correctly specified, in the sense that the fitted values generated under this model are unbiased estimators of the corresponding expected values of the indicator variables on the left hand side of (13). The robust model-based approach to estimating the above prediction variance (see Royall and Cumberland, 1978) can then be used. This leads to the estimator

$$\hat{V} = N^{-2}\left(\sum\nolimits_s \left((w_i - 1)R_i(t)\right)^2 + \sum\nolimits_{U-s} \hat{P}_j(t)(1 - \hat{P}_j(t))\right). \tag{15}$$

where $\hat{P}_i(t) = \hat{\beta}_{0t} + \sum_{k=1}^{p} \hat{\beta}_{kt} I(x_i \le Q_x(\alpha_k))$ and $R_i(t) = I(y_i \le t) - \hat{P}_i(t)$.

4. Application to the Pay Rate Distribution Problem

We return to the problem of estimating the distribution of hourly pay rates using NES data. In section 2 we described methods that allow us to impute the unobserved "complete data" empirical (1). However, we can also estimate this quantity. In particular, setting $n = N$, $s_1 = s$

and $s_2 = U\text{-}s$, we can estimate (1) using the semi-parametric CD-type distribution function estimator (9), with the sample X–values serving as the values of the auxiliary variable. Alternatively, we can compute the X–calibrated HD distribution function estimator (13) of (1). In this section we report results from a simulation study based on data that allows us to evaluate these different approaches. This simulation study uses the 59590 NES respondents who provided data on both Y and X (see Table 2 and Figure 1). A list of the different estimators for which we report simulation results is set out in Table 3.

Table 3 about here.

Since Y and X are supposed to be measuring the same thing, we assume $\mu(x) = \alpha + \beta x$ in both the CD estimator (7) and its semi-parametric version (9). We also assume $\sigma(x)$ is constant in the former. However, since the NES data clearly contain many outliers, we estimate α and β via robust regression using a modified version of the function rlm (Venables and Ripley, 2002) in R (R Development Core Team, 2004).

We simulated two different scenarios representing alternative ways in which hourly pay rate data could become "unavailable". The first, which we refer to below as the "MAR Scenario", is where the probability of a value of Y being unavailable is proportional to the corresponding value of X. In particular, under this scenario, the NES sample data were randomly split into 2 groups, U_1 of size 29590 and U_2 of size 30000 so that Pr(inclusion in U_2) was proportional to X^2. 500 simple random samples s_1 and s_2 each of size 500 were then independently drawn from S_1 and S_2 respectively. Values of Y and X were assumed to be available on s_1, while values of X only were assumed to be available on s_2. The second scenario, which we refer to below as the "Not MAR Scenario", was simulated in exactly the same way as the MAR Scenario, except that Pr(inclusion in U_2) was proportional to Y^2. Table 4 shows the quantiles of the two subpopulations (Y available/Y not available) defined as a consequence. Comparing these values with those in Table 1, we see that the simulated populations are actually less extreme than the reality of the NES data.

Table 4 about here.

In both scenarios the target values of t were the 25 equally spaced values 400, 425, 450, 475, …., 950, 975, 1000. Furthermore, since smaller values of t are more important (reflecting the focus on the lower tail of the pay rate distribution), a weighted CV methodology was used to select the bandwidth coefficient f in (9) and the number of knots f in (13), with the "importance" weights θ_k used in the weighted integrated squared distance criterion ranging in equal decrements from 25 for $t = 400$ to 1 for $t = 1000$.

Tables 5 to 8 show the prediction bias and root mean squared error of each of the estimators defined in Table 3 at each of 25 "target" values of t, where the prediction bias is defined as the average difference between the estimator value and the value of (1) over the 500 simulations, and the root mean squared error is the square root of the average squared difference. Note that separate results are provided for the MAR and NotMAR scenarios.

Tables 5 – 8 about here.

Inspection of the results in Tables 5 – 8 allows one to reach a number of clear conclusions. First, ignoring the information in s_2 and estimating (1) via the "available data" estimator (2) is generally a very poor choice. This estimator is heavily biased under both scenarios. Of the

remaining estimators, all three versions C5, C25 and CCV of the calibrated HD estimator performed very similarly. Not surprisingly, this performance was substantially better in the MAR scenario compared with the NotMAR scenario. The two semi-parametric versions of the CD estimator (L25 and LCV) also performed very similarly, and again were noticeably better in the MAR scenario compared with the NotMAR scenario. They also clearly dominated the calibrated HD estimators in both scenarios. In contrast, the performance of the parametric CD estimator L1 was actually more like that of the substitution estimator SUB rather than like that of the semi-parametric CD estimators. In particular, in the MAR scenario both L1 and SUB had very similar levels of bias and root mean squared error, while in the NotMAR scenario their biases behaved differently, but their levels of root mean squared error were not too dissimilar. In the MAR scenario these estimators exhibited significant bias for medium to large values of t, leading to poor root mean squared error performance, while in the NotMAR scenario they produced the best overall root mean squared error performance. Finally, although essentially unbiased in the MAR scenario, the nearest neighbour estimator NNI was not as efficient as the semi-parametric CD estimators L25 and LCV in this case, while its bias and root mean squared error performance in the NotMAR scenario was on a par with the calibrated HD estimators C5, C25 and CCV.

If we use ">" to denote "performs better than", and bracket similar performing methods, then in our simulation we summarise the relative performances of the different estimators as follows. For the MAR scenario, we have (L25,LCV) > C5 > NNI > (C25,CCV) > (L1,SUB) > Fn, while for the NotMAR scenario this ranking becomes (L1,SUB) > (L25,LCV) > (NNI,C5,C25,CCV) > Fn. Since the MAR scenario is more likely to be the real reason why hourly pay rate data are unavailable, we conclude that, among the different methods considered in our study, using a version of the semi-parametric CD estimator (9) seems the best approach to take when estimating (1).

Although not the primary focus of the study, the simulations were also used to assess the performances of the variance estimators and associated confidence interval estimates generated by the "fixed bandwidth" CD estimator L25 and the two HD estimators C5 and C25 under the MAR scenario. In particular, we computed approximate 95 per cent "2 sigma" confidence intervals for the value of (1) at each value of t in each simulation and then measured the coverage of these intervals over the simulations. These intervals were computed as the estimate value plus or minus twice the square root of the estimated prediction variance. Relative biases of the different variance estimates, and the coverages of their associated confidence intervals are set out in Tables 9 and 10.

The large sample variance estimator (10) of the semi-parametric CD estimator L25 is too numerically intensive to simulate, so we instead simulated its grouped approximation (11), with groups defined by splitting the range of non-sample X-values into G equal sized intervals. We used G = 5, 10 and 25. Unpublished simulations based on random sampling from a number of other "more balanced" data sets used in survey sampling research had indicated that this method works reasonably well with G = 25, typically leading to 2-sigma intervals with coverages exceeding 90%. However, because of the very unbalanced nature of the samples in this application we anticipated a lower value of G would be necessary to get adequate coverage.

In the case of the calibrated HD estimators C5 and C25, we used the robust prediction variance estimator (15), which, as previously noted, is essentially the same as the design-based variance estimator suggested by Harms and Duchesne (2004).

Tables 9 and 10 about here.

The results set out in Tables 9 and 10 are rather surprising, and reinforce the fact that getting good quality measures of the variability of the prediction error is non-trivial given the rather unbalanced sample configurations we observed in our simulations. As expected, both the bias of the grouped variance estimator (11) and its associated confidence interval coverage gets better as G decreases. However, these two measures of performance do not improve at the same rate. In particular, at $G = 5$ the variance estimator (11) is distinctly conservative, with a positive average relative bias of around 15%. Unfortunately, this does not lead to improved coverage. In fact, the same variance estimator leads to confidence intervals with an average coverage of 87%, which is considerably less than the target level of 95%. Furthermore, the situation is reversed when we consider the HD-based variance estimator (15). In the case of C5, we see that this variance estimator has a substantial negative average relative bias of close to 30%, but a much better coverage performance, averaging over 90%. We also note that the variance estimators associated with lager values of G and with increased calibration constraints for the HD estimator are biased much too low to be of any value.

Why do bias and coverage go in opposite directions for (11) and (15)? We have no theoretical answer to this question at present. However, an empirical insight into the reason for the better coverage behaviour of (15) can be obtained from Figure 2, which shows the change in the correlation between the squared prediction error and the value of the variance estimate as t changes. We can immediately see that the variance estimates generated by (15) have a strong positive correlation with the size of the prediction error, leading to wider confidence intervals when errors are larger and hence better coverage. In contrast, there is a much weaker negative correlation between the size of the prediction error and the size of the variance estimate when we use (11) with $G = 5$, implying a slight narrowing of confidence intervals when errors are larger and hence poorer coverage.

Figure 2 about here.

5. Summary and Conclusions

This paper has focussed on the practical problem of deciding between an imputation-based strategy and an estimation-based strategy when adjusting for missing hourly wage rate data in a large national survey of UK employees. Our analysis is largely based on the premise that these data are missing because the relevant employees are not paid by the hour. An alternative "derived" measure of the hourly rate, based on total wages and hours worked data also collected in the survey can be calculated, and we contrast imputation methods based on this derived rate with estimation methods that use this derived rate as an auxiliary. In doing so we note that a number of authors (Rodgers, Brown and Duncan, 1993; Skinner *et. al.*, 2002; Beissel-Durrant, 2003) have observed that such derived rates are typically subject to measurement error and are therefore different from the actual hourly wage rate of an employee at the time of the survey. The basis of our comparison is a simulation study, using both actual and derived wage rate data collected in the survey, which tries to mimic two possible mechanisms for missingness. The MAR scenario assumes that whether or not an hourly wage rate is paid depends entirely on the value of the derived rate (i.e. on salary and hours), while the NotMAR scenario assumes that this missingness actually depends on the value of the actual hourly rate. Under both scenarios the distribution of the derived rates for cases reporting actual rates is very different (shifted to the left) compared with the same distribution for those cases where actual rates are missing. Our aim in both scenarios was to

use the derived wage rate data to predict the value of the sample empirical cumulative distribution function of actual wage rates when there are no missing data.

We draw a number of tentative conclusions from this simulation analysis:

- Ignoring the information in the derived rate data and just estimating the distribution using the reported wage rates leads to highly biased predictions of the "complete data" distribution function.
- Nearest neighbour imputation is (relatively) unbiased but inefficient under MAR and substantially biased under NotMAR.
- Substitution imputation fails for MAR, but works reasonably under NotMAR.
- Under MAR the parametric CD distribution function estimator shows a substantial bias due to model misspecification. However, under NotMAR this estimator performs rather well.
- The semi-parametric version of the CD distribution function estimator works well under MAR. This performance deteriorates under NotMAR.
- Under MAR, the calibrated HD distribution function estimator works reasonably provided only a few calibration constraints are imposed. Increasing the number of constraints generally worsens efficiency. Like the semi-parametric CD estimator, this estimator performs badly under NotMAR.
- Using ordered half sample cross-validation to choose the "bandwidth" parameter f of the semiparametric CD estimator and the HD estimator does not lead to efficiency gains compared with fixing the value of this parameter. On the other hand it also loses little efficiency against fixed f and seems a simple way of choosing this parameter.
- Variance estimation, and associated confidence interval estimation, is difficult in the very unbalanced situations explored in the simulation study. We obtained conflicting results under MAR, with reasonable variance estimation for the semi-parametric CD estimator but poor confidence interval estimation, compared with poor variance estimation for the HD estimator but reasonable confidence interval estimation. This remains a topic for further research.

Bearing in mind that a MAR-type mechanism for missing hourly wage rate data seems more likely than a NotMAR mechanism, our overall conclusion is that an estimation approach using the robust semi-parametric CD estimator seems the best way of using the information in the derived wage rate data to recover the "complete data" sample empirical distribution function of actual wage rates. The other methods of estimation we considered were not as efficient overall and imputation-based methods seemed generally inferior to estimation-based methods.

The sample size used in the simulation study (n = 1000) was much smaller than the actual sample size (n = 162,843) in the 2002 NES. In particular, data obtained in this survey allow derived rates to be calculated for n = 153,611 employees from employers who could be classified to one of the industry groups defined by the Standard Industry Classification used in ONS surveys. This sample is made up of 71,382 employees for whom values for both actual and derived rates are available and 82,229 employees for whom only values of derived rates are available. One could therefore ask whether there is any discernable difference between the different methods discussed above when these "full" NES sample data are used. Figure 3 shows the estimated cumulative distributions of actual hourly wage rates for all UK industries generated by the different methods using these data. Given the interest in the

distribution of low wage rates, the plot is restricted to £6 an hour or less, with each distribution displayed there equal to the weighted average of the corresponding estimated distributions fitted within industry groups. Note the significant differences between the distribution estimates shown there, even at this high level of aggregation. For example, the estimated proportions of employees with hourly wage rates less than £6 are 46.2%, 31.5%, 35.4%, 33.8% and 36.1% using Fn, SUB, NNI, L25 and C5 respectively. Adjusting the RMSE results in Table 6 to allow for the much larger sample size applicable here, we see that the RMSEs associated with these estimates can be expected to vary from 1.4% (Fn) to 0.1% (L25). On the basis of our simulation results we would argue that the estimate based on just the employees with reported values of actual hourly wage rates is clearly biased towards too many low paid employees, while the substitution estimate is biased towards too many high paid employees. Our preferred estimate is the one defined by L25.

Figure 3 about here.

References

Beissel-Durrant, G. (2003). Correcting for measurement error when estimating pay distributions from household survey data. *Unpublished Ph.D. Thesis*, School of Social Sciences, University of Southampton.

Chambers, R.L. and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika* **73**, 597 - 604.

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics* **12**, 3 - 32.

Deville, J. C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376 - 382.

Dorfman, A.H. and Hall, P. (1993). Estimators of the finite population distribution function using nonparametric regression. *Annals of Statistics* **21**, 1452 – 1475.

Dunstan, R. and Chambers, R.L. (1989). Estimating distribution functions from survey data with limited benchmark information. *Australian Journal of Statistics* **31**, 1 - 11.

Harms, T. and Duchesne, P. (2004). Calibration estimation for quantiles. *Proceedings of the Joint Statistical Meetings of the American Statistical Association*, Toronto, August 2004.

Lombardia *et al*. (2004). Estimation of a finite population distribution function based on a linear model with unknown heteroscedastic errors. *Working Paper*, Department of Statistics, University of Santiago de Compostela.

R Development Core Team (2004). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org.

Rodgers, W.L., Brown, C. and Duncan, G.J. (1993). Errors in survey reports of earnings hours worked and hourly pay. *Journal of the American Statistical Association* **88**, 1208 – 1218.

Royall, R.M. and Cumberland, W.G. (1978). Variance estimation in finite population sampling. *Journal of the American Statistical Association* **73**, 351 - 358.

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.

Venables, W.N. and Ripley, B.D. (2002) *Modern Applied Statistics with S*. Fourth edition. New York: Springer.

ANNEX: Estimating the Large Sample Prediction Variance of the Semi-Parametric Estimator

From (9) we can write this prediction variance as $N^2 Var\left(\hat{F}_{CDL}(t) - F_N(t)\right) = V_s + V_{U-s}$ where

$$V_s = Var\left(\sum_{U-s} \hat{P}(t, x_j)\right) = \sum_{j \in U-s} \sum_{k \in U-s} Cov\left(\hat{P}(t, x_j), \hat{P}(t, x_k)\right)$$

and

$$V_{U-s} = \sum_{U-s} Var\left(I(y_j \le t)\right).$$

Here $\hat{P}(t, x_j)$ and w_{ij}^* are defined following (10). In order to estimate V_s, we note that since $y_i = \mu(x_i) + \sigma(x_i)\varepsilon_i$, in large samples we can replace $\hat{P}(t, x_j)$ by $P_j(t) = \sum_{i \in s} w_{ij}^* I\left(\varepsilon_i \le u_{ij}(t)\right)$ where $u_{ij}(t) = \sigma^{-1}(x_i)(t - \mu(x_j))$. Then

$$Cov\left(P_j(t), P_k(t)\right) = \sum_{i \in s} w_{ij}^* w_{ik}^* Cov\left(I\left(\varepsilon_i \le u_{ij}(t)\right), I\left(\varepsilon_i \le u_{ik}(t)\right)\right) = \Psi_{jk}^{(1)} - \Psi_{jk}^{(2)}$$

with

$$\begin{aligned}\Psi_{jk}^{(1)} &= \sum_{i \in s} w_{ij}^* w_{ik}^* \Pr\left(\varepsilon_i \le \min(u_{ij}(t), u_{ik}(t))\right) \\ &= \sum_{i \in s} w_{ij}^* w_{ik}^* \Pr\left(y_i - \mu(x_i) \le t - \max(\mu(x_j), \mu(x_k))\right)\end{aligned}$$

and

$$\begin{aligned}\Psi_{jk}^{(2)} &= \sum_{i \in s} w_{ij}^* w_{ik}^* \Pr\left(\varepsilon_i \le u_{ij}(t)\right) \Pr\left(\varepsilon_i \le u_{ik}(t)\right) \\ &= \sum_{i \in s} w_{ij}^* w_{ik}^* \left\{\Pr\left(y_i - \mu(x_i) \le t - \mu(x_j)\right) \Pr\left(y_i - \mu(x_i) \le t - \mu(x_k)\right)\right\}.\end{aligned}$$

Put $P_i(u) = \Pr(y_i - \mu(x_i) \le u)$ and assume that the non-sample units are labelled from 1 to *N-n* in the same order as their mean values. That is, $\mu(x_j) < \mu(x_k)$ when $j < k$. We can then write

$$
\begin{aligned}
V_s &= \sum_{j=1}^{N-n}\sum_{k=1}^{N-n}\Big\{\sum\nolimits_{i\in s} w_{ij}^{*}w_{ik}^{*}P_i(t-\max(\mu(x_j),\mu(x_k))-\sum\nolimits_{i\in s} w_{ij}^{*}w_{ik}^{*}P_i(t-\mu(x_j))P_i(t-\mu(x_k))\Big\} \\
&= \sum\nolimits_{i\in s}\left(\sum_{j=1}^{N-n}\sum_{k=1}^{N-n} w_{ij}^{*}w_{ik}^{*}\left\{P_i(t-\max(\mu(x_j),\mu(x_k))-P_i(t-\mu(x_j))P_i(t-\mu(x_k))\right\}\right) \\
&= \sum\nolimits_{i\in s}\left(\sum_{j=1}^{N-n}\sum_{k=1}^{N-n} w_{ij}^{*}w_{ik}^{*}\left\{P_i(t-\max(\mu(x_j),\mu(x_k))\Big(1-P_i(t-\min(\mu(x_j),\mu(x_k))\Big)\right\}\right) \\
&= \sum\nolimits_{i\in s}\left\{\sum_{j=1}^{N-n} w_{ij}^{*}\Big(1-P_i(t-\mu(x_j))\Big)\right\}\left\{\sum_{j=1}^{N-n} w_{ij}^{*}P_i(t-\mu(x_j))\right\} \\
&\qquad -\sum_{j=1}^{N-n}\sum_{k=1}^{j} w_{ij}^{*}w_{ik}^{*}\left\{P_i(t-\mu(x_k))-P_i(t-\mu(x_j))\right\}
\end{aligned}
$$

The estimator $\hat{V}_s$ defined after (10) follows immediately when we replace $P_i(u)$ by $\hat{P}_{si}(u)$. Estimating V_{U-s} on the other hand is more straightforward. Since

$$V_{U-s} = \sum\nolimits_{U-s}\Pr(y_j \le t)\Big(1-\Pr(y_j \le t)\Big)$$

all we need to do is to replace $\Pr(y_j \le t)$ by its obvious estimator $\hat{P}(t, x_j)$.

Table 1 Distribution of NES data for 2002 based on the total sample of 162843 employees, of whom 75850 provided hourly pay rate data (Y). All values are in pence.

		Quantiles of distribution		
Y available?		25%	50%	75%
Yes	Y	482	597	843
	X	492	634	892
No	X	717	1014	1491

Table 2 Distributions of Y and X for the n = 59590 employees that providing these values and with $300 \le Y \le 2000$ and $300 \le X \le 3000$. All values are rounded to the nearest five pence.

Quantile	Y	X
100.0%	1995	2955
90.0%	1120	1190
75.0%	820	870
50.0%	600	635
25.0%	495	500
10.0%	435	440
0.0%	300	300

Table 3 Labels and definitions for the estimators considered in the simulation study.

Label	Definition
Fn	Available data estimator (2).
SUB	Substitution imputation estimator (3).
NNI	Limiting form of nearest neighbour imputation estimator (5), with imputed Y-value drawn randomly from three nearest neighbours.
L1	Parametric CD estimator (7) under $\mu(x) = \alpha + \beta x$ and $\sigma(x)$ = constant.
L25	Semiparametric CD estimator (9) with $\mu(x) = \alpha + \beta x$ and with bandwidth coefficient f = 25.
LCV	Semiparametric CD estimator (9) with $\mu(x) = \alpha + \beta x$ and with bandwidth coefficient f chosen via weighted cross validation from f = 5,10,15, …, 45,50.
C5	Calibrated HD estimator with calibration constraints defined by $\alpha_k = kf^{-1}; k = 1,\ldots, f-1$ with f = 5.
C25	Calibrated HD estimator with calibration constraints defined by $\alpha_k = kf^{-1}; k = 1,\ldots, f-1$ with f = 25.
CCV	Calibrated HD estimator with calibration constraints defined by $\alpha_k = kf^{-1}; k = 1,\ldots, f-1$ with f chosen via weighted cross validation from f = 5,10,15, …, 45,50.

Table 4 Quantiles of simulated populations

Y available?		Quantiles of distribution				
		10%	25%	50%	75%	90%
				MAR Scenario		
Yes	*Y*	420	465	530	650	810
	X	420	465	545	670	835
No	*Y*	470	550	725	1000	1350
	X	485	589	790	1065	1435
				Not MAR Scenario		
Yes	*Y*	420	460	525	640	790
	X	420	470	550	690	865
No	*Y*	470	550	740	1010	1365
	X	475	575	770	1050	1420

Table 5 Values of prediction bias ($\times 10^4$) for MAR scenario.

t	Fn	SUB	NNI	L1	L25	LCV	C5	C25	CCV
400	34	3	0	135	39	41	5	5	5
425	434	-40	3	63	-14	-11	24	15	17
450	726	-80	2	7	-46	-41	22	28	29
475	950	-95	-1	4	-15	-13	16	43	44
500	1259	-178	-2	-111	-79	-82	28	58	60
525	1437	-215	9	-134	-63	-65	52	87	90
550	1577	-269	-3	-205	-93	-96	45	94	98
575	1665	-273	3	-209	-61	-70	52	106	109
600	1724	-300	-13	-242	-67	-81	63	97	103
625	1753	-330	-24	-262	-67	-84	63	97	102
650	1768	-340	-32	-275	-70	-85	53	94	100
675	1756	-373	-34	-307	-96	-109	44	101	107
700	1747	-378	-29	-318	-109	-118	46	113	117
725	1721	-362	-35	-297	-89	-97	59	116	121
750	1695	-346	-23	-282	-82	-91	97	132	140
775	1650	-310	-27	-248	-59	-69	100	131	139
800	1588	-311	-30	-255	-88	-96	98	130	140
825	1538	-288	-32	-233	-79	-87	101	133	141
850	1474	-285	-36	-233	-93	-101	97	132	141
875	1421	-277	-35	-217	-89	-96	97	129	138
900	1362	-266	-16	-210	-93	-99	111	144	155
925	1293	-270	-2	-215	-105	-110	147	164	173
950	1228	-254	0	-207	-100	-107	185	163	177
975	1166	-248	11	-198	-96	-106	247	168	187
1000	1087	-253	26	-209	-115	-127	334	178	203

Table 6 Values of prediction root mean squared error ($\times 10^4$) for MAR scenario.

t	Fn	SUB	NNI	L1	L25	LCV	C5	C25	CCV
400	41	**14**	54	137	42	49	18	23	23
425	443	50	139	72	**36**	39	56	48	51
450	735	89	181	**43**	62	60	63	75	74
475	958	105	196	**44**	50	49	64	93	93
500	1266	186	219	122	98	101	**84**	120	120
525	1444	223	233	146	**91**	98	105	152	155
550	1584	275	242	213	118	128	**107**	169	171
575	1672	279	245	217	**97**	111	114	187	185
600	1730	306	247	250	**102**	119	124	187	187
625	1759	336	253	269	**106**	123	133	200	198
650	1774	346	253	283	**110**	123	129	204	200
675	1762	379	261	315	**133**	144	139	221	221
700	1753	384	263	325	143	151	**141**	232	230
725	1726	368	273	305	**130**	135	152	241	238
750	1700	352	265	290	**127**	132	176	252	252
775	1655	315	255	256	**111**	116	174	252	249
800	1593	317	250	263	**128**	134	174	255	254
825	1543	294	256	242	**120**	125	179	263	258
850	1479	292	258	241	**128**	133	178	266	260
875	1426	283	253	226	**124**	130	177	261	257
900	1367	273	259	219	**127**	133	191	277	274
925	1298	276	259	225	**137**	141	221	298	291
950	1233	260	261	216	**133**	138	256	301	295
975	1171	254	262	207	**130**	138	312	306	304
1000	1092	260	267	217	**144**	154	390	314	322

Table 7 Values of prediction bias ($\times 10^4$) for NotMAR scenario.

t	Fn	SUB	NNI	L1	L25	LCV	C5	C25	CCV
400	43	11	12	216	62	61	20	21	22
425	465	-2	89	142	39	43	113	92	95
450	760	-18	142	106	53	59	155	152	153
475	981	-20	182	122	135	130	188	202	202
500	1313	-89	252	40	119	110	274	269	274
525	1510	-104	323	50	185	182	354	345	350
550	1661	-139	385	11	204	201	405	415	417
575	1758	-136	431	39	279	268	449	467	465
600	1839	-139	491	41	315	299	517	521	522
625	1880	-155	526	53	351	328	563	564	563
650	1890	-162	547	59	371	348	577	583	582
675	1890	-190	586	46	367	343	607	623	624
700	1881	-199	620	39	358	340	631	651	652
725	1851	-187	627	66	383	365	646	661	664
750	1813	-178	636	86	388	365	672	672	674
775	1779	-148	634	129	408	383	687	671	675
800	1721	-156	644	123	375	352	697	678	683
825	1665	-141	642	143	369	349	692	676	681
850	1594	-141	638	142	345	323	674	665	667
875	1531	-138	633	158	338	313	658	660	663
900	1450	-140	616	147	306	283	632	643	645
925	1371	-152	610	133	274	252	635	638	640
950	1296	-145	586	130	254	233	627	612	613
975	1235	-141	577	138	242	219	653	603	608
1000	1152	-152	565	117	203	181	677	587	599

Table 8 Values of prediction root mean squared error ($\times 10^4$) for NotMAR scenario.

t	Fn	SUB	NNI	L1	L25	LCV	C5	C25	CCV
400	50	18	31	217	65	68	32	44	46
425	473	31	118	147	54	58	128	118	119
450	768	43	178	114	72	76	171	180	179
475	989	48	219	129	146	141	205	230	228
500	1320	101	290	63	136	128	292	300	302
525	1516	115	360	73	200	198	371	374	377
550	1667	149	421	59	222	220	424	444	445
575	1763	148	466	74	295	286	465	494	492
600	1844	152	529	81	332	317	536	551	549
625	1886	167	565	92	367	346	583	594	590
650	1895	173	587	97	387	366	596	613	608
675	1894	201	626	94	383	361	626	653	652
700	1886	209	661	94	376	359	649	680	680
725	1855	197	668	113	400	383	664	690	693
750	1817	188	679	126	405	384	691	703	704
775	1784	160	678	161	424	401	707	702	706
800	1726	167	685	157	392	373	716	708	712
825	1669	153	684	175	387	370	711	707	711
850	1598	153	678	176	363	346	693	696	697
875	1536	149	672	187	356	337	676	691	690
900	1454	152	657	179	326	310	651	676	674
925	1376	163	651	167	295	281	653	670	669
950	1301	155	629	165	277	263	646	646	645
975	1239	152	621	172	266	252	671	637	641
1000	1156	161	610	153	230	219	693	622	633

Table 9 Ratio of average of variance estimates to actual mean squared error under the MAR scenario. Variance estimates for L25 are defined by (11) and those for C5, C25 are defined by (15).

t	L25			C5	C25
	$G = 5$	$G = 10$	$G = 25$		
400	0.5820	0.7735	0.7142	1.0557	0.7341
425	1.2505	1.4939	1.2845	0.7485	0.6729
450	0.6568	0.7125	0.5236	0.9093	0.5170
475	1.5716	1.4995	0.9214	0.9863	0.4356
500	0.5913	0.4428	0.2891	0.9060	0.3544
525	0.9866	0.4932	0.3868	0.7295	0.2775
550	0.8054	0.3100	0.2627	0.8171	0.2649
575	1.4200	0.5222	0.4311	0.8314	0.2410
600	1.4110	0.5741	0.4494	0.8264	0.2581
625	1.2153	0.6393	0.4565	0.7801	0.2418
650	0.9868	0.6657	0.4505	0.8473	0.2400
675	0.6542	0.5031	0.3297	0.8097	0.2261
700	0.5830	0.4562	0.3015	0.8354	0.2099
725	0.7705	0.5516	0.3733	0.7649	0.2032
750	0.8834	0.5539	0.4020	0.6207	0.1883
775	1.2633	0.6978	0.5369	0.6674	0.1900
800	1.0312	0.5400	0.4085	0.6730	0.1864
825	1.3324	0.6561	0.4663	0.6559	0.1795
850	1.2966	0.6087	0.4057	0.6928	0.1822
875	1.5120	0.6572	0.4280	0.7170	0.1849
900	1.5987	0.6393	0.4077	0.6580	0.1746
925	1.4857	0.5541	0.3433	0.5594	0.1612
950	1.6568	0.5849	0.3523	0.4574	0.1548
975	1.7695	0.5802	0.3634	0.3403	0.1521
1000	1.4540	0.4312	0.2912	0.2293	0.1459
Average	1.1508	0.6456	0.4632	0.7248	0.2711

Table 10 Actual coverages of nominal 95% "2-sigma" confidence intervals under the MAR scenario. Variance estimates for L25 are defined by (11) and those for C5, C25 are defined by (15).

t	L25			C5	C25
	$G = 5$	$G = 10$	$G = 25$		
400	0.634	0.830	0.926	0.948	0.930
425	0.874	0.946	0.948	0.924	0.916
450	0.818	0.888	0.786	0.950	0.858
475	0.952	0.966	0.930	0.962	0.830
500	0.808	0.778	0.680	0.948	0.782
525	0.930	0.788	0.740	0.938	0.716
550	0.918	0.674	0.622	0.940	0.714
575	0.968	0.810	0.790	0.954	0.694
600	0.942	0.822	0.786	0.950	0.688
625	0.914	0.846	0.784	0.938	0.690
650	0.872	0.864	0.774	0.958	0.682
675	0.784	0.804	0.726	0.942	0.662
700	0.754	0.784	0.686	0.922	0.622
725	0.802	0.816	0.720	0.926	0.612
750	0.836	0.834	0.768	0.904	0.594
775	0.874	0.862	0.824	0.910	0.572
800	0.868	0.810	0.772	0.908	0.570
825	0.894	0.854	0.790	0.902	0.550
850	0.882	0.820	0.744	0.924	0.520
875	0.908	0.848	0.750	0.922	0.520
900	0.926	0.838	0.742	0.928	0.512
925	0.938	0.822	0.694	0.884	0.516
950	0.940	0.830	0.698	0.846	0.498
975	0.920	0.796	0.712	0.760	0.496
1000	0.880	0.726	0.612	0.578	0.490
Average	0.873	0.826	0.760	0.907	0.649

Figure 1 Scatterplot of observed hourly pay rate (Y) versus derived hourly pay rate (X) for the 59590 employees in the NES sample with $300 \leq Y \leq 2000$ and $300 \leq X \leq 3000$.

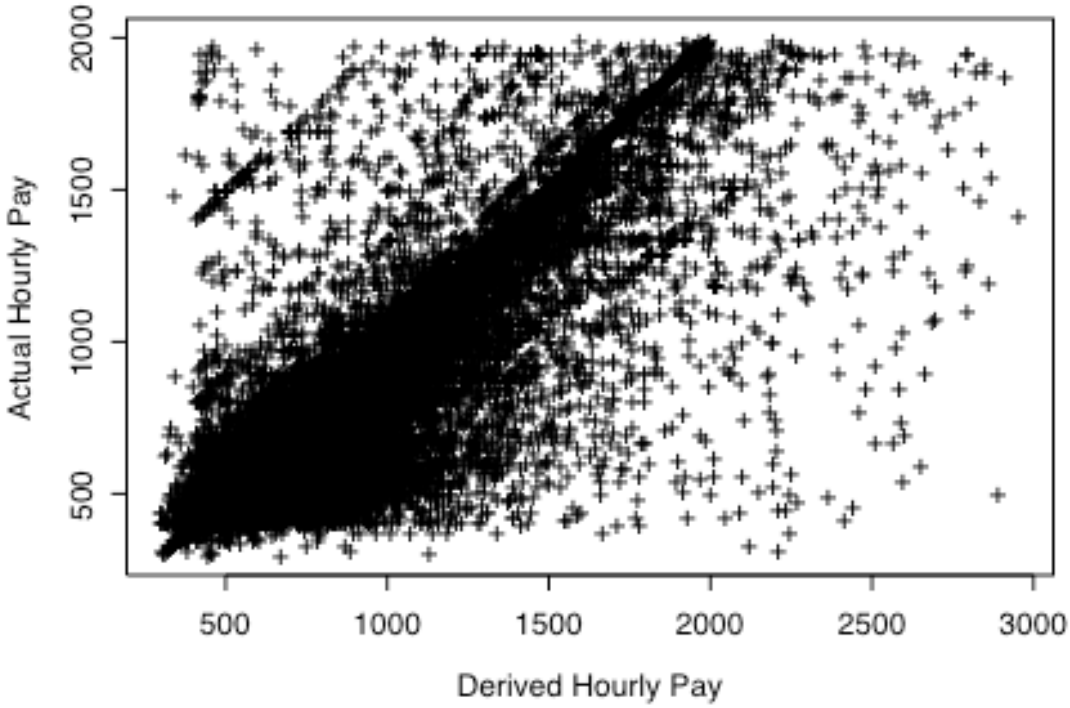

Figure 2 Change in correlation between squared estimation error and value of variance estimate as t changes. Solid line is L25, with variance estimated using (11) and $G = 5$; Dashed line is C5 with variance estimated via (15).

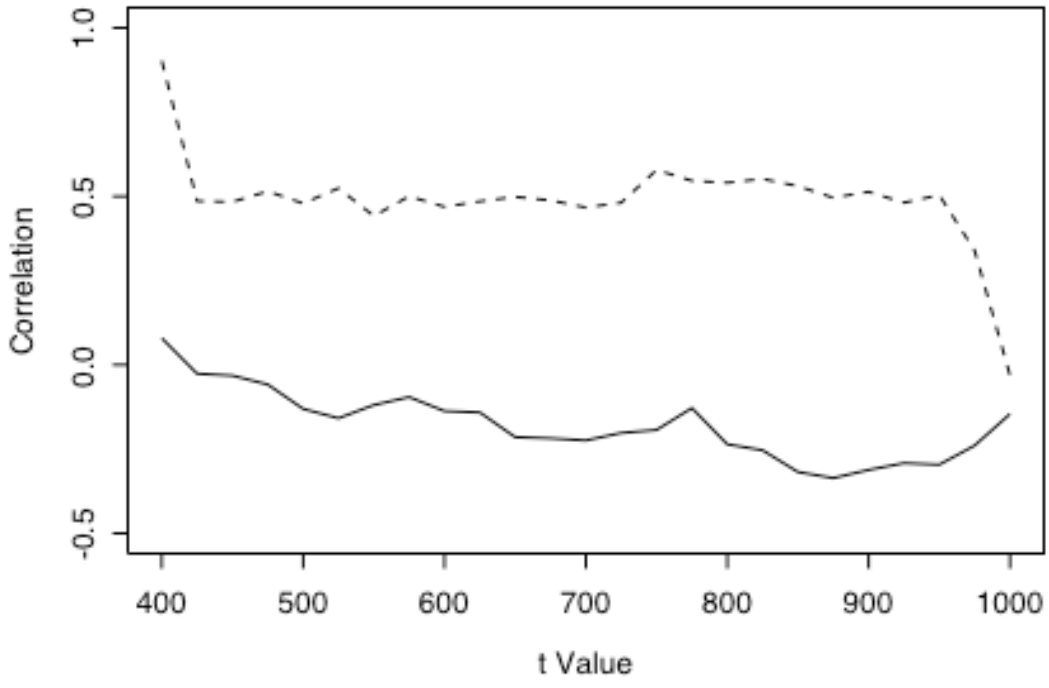

Figure 3 Different estimated hourly wage rate cumulative distributions based on NES data from 2002. Solid line is based on available sample data for hourly wage rates; short dashed line is based on substituting derived rates when actual rates are missing; dotted line is asymptotic version of near neighbour imputation estimator; dash-dot line is semi-parametric estimator L25 and long dashed line is calibrated spline estimator C5.