THE MEASUREMENT OF LOW PAY IN THE UK LABOUR FORCE SURVEY

CHRIS SKINNER, NIGEL STUTTARD, GABRIELE BEISSEL-DURRANT,
JAMES JENKINS

ABSTRACT

Consideration of the National Minimum Wage requires estimates of the distribution of hourly pay. The UK Labour Force Survey (LFS) is a key source of such estimates. The approach most frequently adopted by researchers has been to measure hourly earnings from several questions on pay and hours. The Office for National Statistics is now applying a new approach, based on an alternative more direct measurement introduced in March 1999.

These two measures do not produce identical values and this paper investigates sources of discrepancies and concludes that the new variable is more accurate. The difficulty with using the new variable is that it is only available on a subset of respondents. An approach is developed in which missing values of the new variable are replaced by imputed values. The assumptions underlying this imputation approach and results of applying it to LFS data are presented. The relation to weighting approaches is also discussed.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M03/04

University
of Southampton

The Measurement of Low Pay in the U.K. Labour Force Survey

Chris Skinner, University of Southampton

Nigel Stuttard, Office for National Statistics

Gabriele Beissel-Durrant, University of Southampton

James Jenkins, Office for National Statistics

Abstract

Consideration of the National Minimum Wage requires estimates of the distribution of hourly pay. The UK Labour Force Survey (LFS) is a key source of such estimates. The approach most frequently adopted by researchers has been to measure hourly earnings from several questions on pay and hours. The Office for National Statistics is now applying a new approach, based on an alternative more direct measurement introduced in March 1999. These two measures do not produce identical values and this paper investigates sources of discrepancies and concludes that the new variable is more accurate. The difficulty with using the new variable is that it is only available on a subset of respondents. An approach is developed in which missing values of the new variable are replaced by imputed values. The assumptions underlying this imputation approach and results of applying it to LFS data are presented. The relation to weighting approaches is also discussed.

JEL classification: C42 (survey methods)

Key words: distribution function; hourly pay; imputation; measurement error; missing data; weighting.

Acknowledgements

The authors are grateful to the Editors and two referees for comments which led to significant improvements to the paper and to Dick Heasman for valuable input.

Corresponding author:

Nigel Stuttard

Office for National Statistics, B3 02, 1 Drummond Gate, London SW1V 2QQ

Email: nigel.stuttard@ons.gov.uk

Telephone: 020 7533 6167

1. Introduction

Estimates of the distribution of hourly pay are needed to study the effects of the introduction of the National Minimum Wage (NMW) and changes in minimum rates, as well as to inform judgements about how these rates might be changed. The principal data sources used by the Office for National Statistics (ONS) to estimate this distribution are the Labour Force Survey (LFS) and the New Earnings Survey (NES). Both sources have their strengths and limitations for low pay estimates. The LFS is a household survey, which has good coverage, is conducted quarterly and has a wide range of variables, which may be used for analysis. It suffers, however, from the problem of measurement error which, as discussed in this paper, can lead to serious overestimation of the proportions of low paid, especially at the extreme of the distribution. The NES is an employer survey, which has a larger sample size than the LFS and includes hourly pay information, which is considered accurate, being derived from pay rolls. The NES is only conducted annually, however, and currently suffers from under-coverage of some kinds of low-paid employees, especially those earning below the PAYE (Pay-as-you-earn) tax threshold. As a result, large discrepancies have in the past occurred between unadjusted LFS and NES estimates of the proportions earning below low pay thresholds, with LFS estimates consistently higher. For example, unadjusted estimates of the proportions earning below £2.50 per hour in 1997 were 4.2% from the LFS versus 1.4% from the NES (Wilkinson, 1998). Since then, various steps have been taken to improve LFS and NES estimates, both by introducing a new hourly rate variable into the LFS in March 1999 and by changing estimation methods, such as introducing weighting for NES non-response. These changes have tended to reduce discrepancies between LFS and NES estimates. For example, estimates of the proportions earning below NMW rates in 2000 were 1.4% from the LFS versus 1.0% from the NES (Stuttard and Jenkins, 2001). ONS combines the LFS and NES estimates into a central estimate and these estimates suggest that the introduction of the NMW has led to a sharp fall in the number of jobs with pay below NMW rates but that the distribution above this threshold has been largely unaffected (Stuttard and Jenkins, 2001). Dickens and Manning (2002) draw a similar conclusion.

We focus in this paper on the problem of measurement error for low pay estimates from the LFS and on the recent development of LFS estimation methods to make use of the new hourly rate variable (see the Appendix for an outline of LFS methodology and Stuttard and Jenkins (2001) on the use of the NES for low pay estimates).

There is extensive evidence of measurement error in household survey data on earnings, hours worked and hourly pay, for example from validation studies for the U.S. Current Population Survey (Mellow and Sider, 1983; Bound and Krueger, 1991; Bollinger,1998) and the U.S. Panel Study of Income Dynamics (Duncan and Hill, 1985; Rodgers et al., 1993; Bound et al., 1994). See Moore et al. (2000) for a review. Evidence of measurement error in LFS data will be presented in section 2.

The principal concern we shall have is that such error may bias estimation of the proportions earning below low pay thresholds, such as the NMW. We shall be interested in proportions derived from cumulative distribution functions of the form:

$$F_D(y) = \sum_{i \in D} I(y_i \leq y) / N_D, \qquad (1)$$

where y_i is the hourly rate of pay for job i, y is a specified pay rate such as the NMW, I(A) is the indicator function (=1 if A is true and 0 otherwise), D is a specified set of jobs of interest, e.g. all jobs of men aged 18-21 in a specified region, and N_D is the number of jobs in D. Thus, $F_D(y)$ is the proportion of jobs in D with hourly pay no greater than y. Theoretical arguments (Chesher, 1991; Fuller, 1995) show that measurement error in y_i may lead to overestimation of proportions in the distribution's lower tail.

The established way of measuring hourly pay in the LFS has been to divide gross pay received by usual weekly hours worked. We refer to this as the *derived* hourly pay variable. The earnings and hours questions upon which this variable is based have, however, been designed to meet needs other than the measurement of hourly pay. In order to address this objective better, in the context of the introduction of the NMW, a new variable was introduced into the LFS in March 1999, measuring hourly rate directly. We refer to this as the *direct* hourly rate variable. Although this variable appears to improve greatly on the derived variable as a measure of hourly pay, it suffers from only being available for a subset of respondents. To address this problem, an approach is described in Section 3 in which values of hourly pay are imputed for cases where the direct variable is missing. An alternative weighting approach, proposed by Dickens and Manning (2002), will be considered in section 4.

The LFS collects data not only on main jobs but also on second jobs. These make up only about 3.5% of all jobs, but a higher proportion of low paid jobs. For example, in spring 1999, the proportion of second jobs paid less than NMW rates was estimated to be about ten times greater than for main jobs. As a result, second jobs have a non-negligible effect on low pay estimates for all jobs. The direct hourly rate variable is, however, not collected in the LFS for second jobs. All that is available is a derived variable and this may be subject to the same kinds of measurement errors as for main jobs. ONS is giving further consideration to how estimates for second jobs might be improved, but in this paper we shall restrict attention to main jobs. In this case the units i in the definition in (1) may be considered as either (main) jobs or employees. The basic estimator of the distribution function $F_D(y)$ is then given by

$$\hat{F}_D(y) = \sum_{S_D} w_i I(y_i \le y) / \sum_{S_D} w_i \tag{2}$$

where s_D is that part of the LFS sample falling into domain D and w_i is the survey weight for employee i (see Appendix).

The main parts of this paper consist of an investigation of measurement error in the LFS in section 2 and the development of an imputation approach in section 3. The relation between the imputation approach and weighted estimation is discussed in section 4, some directions of further research are outlined in section 5 and conclusions are summarised in section 6.

2. Measurement of Hourly Pay in the LFS

In this section we consider the nature and extent of measurement error in the two LFS hourly pay variables. Measurement error is defined as the difference between the recorded value of the variable and the value of the ‘true variable’, which we should ideally like to measure. This variable is taken to be the *basic* pay rate, i.e. before any overtime, bonuses or discretionary additions, the *gross* rate, i.e. before deductions such as tax, and the *current* rate, i.e. applying at a specified date. If a fixed (basic) hourly rate is specified for the job then this defines the rate. Otherwise, the rate is the ratio of (basic) pay and (basic) hours, as specified in a job contract. In this way the hourly pay variable is intended to correspond broadly to the pay definition in NMW legislation. Note, however, that estimates of $F_D(y)$ for y=NMW cannot necessarily be used as a measure of non-compliance with the legislation, because it is not possible to discern from the LFS whether an individual is eligible for minimum wage rates and hence to specify D to exclude ineligible individuals. Examples of exceptional cases where the minimum rates do not apply are apprentices and those undergoing training, who are exempt from the minimum wage rate or are entitled to lower rates and employees who receive free accommodation, for whom employers are entitled to offset hourly rates by up to 50p per hour.

2.1. Derived Hourly Pay Variable

The variable traditionally used to measure gross hourly pay in a main job is derived by dividing gross weekly earnings by usual hours worked:

Derived hourly pay variable = GRSSWK / (BUSHR + POTHR),

where GRSSWK is gross weekly earnings from main job, BUSHR is basic usual weekly hours in main job and POTHR is usual weekly paid overtime hours in main job. The variable GRSSWK is itself derived from answers to the following two questions:

What was your gross pay, that is your pay before any deductions, the last time you were paid?

What period did this cover? (GRSPRD)

In order to consider the nature of measurement error, we distinguish two sources of error. First, even if a respondent answers questions 'correctly', *definitional error* arises if the value of the pay rate calculated from these correct answers differs from the true value of interest. Second, respondents may not answer questions correctly, leading to what we call *reporting error.* This is a rough distinction since the term 'correctly' is not entirely well-defined, for example, the correct answer to the question 'what are the usual hours you work?' may not be clear for someone who works irregular hours. There are at least four potential sources of definitional error:

i) The derived variable includes all pay, not just basic pay, so that overtime, bonuses and other additional sources of pay, which are often at higher rates of pay, could lead to positive errors.

ii) The numerator of the derived variable refers to actual earnings whereas the denominator refers to usual hours. Thus, even if all hours are paid at the same rate, the derived variable may not equal this rate if the actual hours worked differs from the usual hours during the pay period. This could lead to positive or negative errors. Moreover, the usual hours may exceed the contracted hours because of 'unpaid overtime', leading to negative errors.

iii) Since the respondent is offered a fixed set of alternative pay periods for the GRSPRD question, some approximation of the true period may occur. Most respondents choose one week, four weeks (or one month) or one year. The worst errors seem likely to be for respondents who select the option "less than one week", when their pay period is assumed to be half a week and GRSSWK is calculated by doubling the reported gross pay. Either a positive or negative error may then arise.

iv) The derived variable refers to last pay received, and current rates could have increased since then, for example to comply with changes in the NMW rates. This may be more of a problem for monthly than for weekly or annual reporters, since those who report for a monthly period may report pay according to rates up to one month old, whereas weekly reporters will tend to refer to a more recent period and annual reporters are expected to refer to their current annual gross pay rate. Errors from this source seem likely to be negative.

There are also several potential sources of reporting error:

v) Just under a third of information collected in the LFS is supplied by proxy by other household members. Information on earnings and hours worked collected in this way is known to be of poorer quality than information collected from personal respondents. Previous research conducted by ONS (Wilkinson,1998) found that, where the proxy information was supplied by the spouse or partner, hours worked tended to be overstated by between 2% and 5% and, where the information was supplied by another adult member of the household, both weekly earnings and hours worked tended to be understated, resulting in hourly earnings being understated by between 6% and 12%.

vi) There is still potential for reporting error when the information is supplied by personal respondents. The figure supplied for gross pay may be rounded or approximated. Although respondents are encouraged to refer to their pay-slip when answering the pay questions, there is no compulsion and many answer without reference to any documentary support. The data on usual hours worked are also affected by respondents giving rounded or approximate answers. In addition, respondents who work irregular hours will find this question difficult to answer.

2.2. Direct Hourly Rate Variable

Two new questions were introduced in March 1999. The first, HOURLY, asks "are you paid a fixed hourly rate?". Respondents who answer "yes" are then asked HRRATE, "what is your (basic) hourly rate?". This defines the direct variable. Initially, the question HOURLY was addressed only to respondents whose pay period is weekly or fortnightly, or who report their pay as a lump sum or do not know their pay period. In March-May 99 this resulted in 4,723 valid responses (unweighted cases) to HRRATE, compared to 17,615 valid responses to the derived variable. From March-May 2000, the question HOURLY was extended to all earnings respondents, resulting in 7,176 valid responses to HRRATE. Nevertheless, the subsample of jobs for which the direct variable is recorded represents less than half the jobs for which the derived variable is available and consists of a selective subsample, since those employees who are paid a fixed hourly rate tend generally to be lower paid than other employees. Some indication of the degree of selectivity is shown in Table 1 which compares various (weighted) summary measures of the distribution of the derived variable for all cases with those where the direct variable is reported. We see, for example, that 50% of all jobs have derived pay rates over £6.67 per hour but that less than 25% of those jobs for which the direct variable is reported have such high pay rates.

[Table 1 about here]

The direct variable suffers from none of the sources of definitional error for the derived variable. In particular, HRRATE refers to a current rate, not the last pay received. There is of course still the potential for respondent error if the respondent, or proxy respondent, forgets or is unaware of the precise hourly rate at the time of the interview. Nevertheless, feedback from the LFS pilot survey, that the questions HOURLY and HRRATE tended to be well understood, suggests that respondent error may be low.

A further reason why the direct variable may be subject to less measurement error than the derived variable is that respondents answering “yes” to HOURLY are not pressed to respond to HRRATE if they indicate that they do not know the rate. In contrast, respondents who express difficulty in providing the gross earnings figure for the derived variable are encouraged to provide an approximate figure rather than no response. Evidence that the direct variable data is less prone to measurement error by ‘guessing’ may be obtained from comparing proxy response rates, since proxy respondents are less likely to know the hourly rate. There is a much higher proportion of proxy respondents among those answering “yes” to HOURLY but not providing a response to HRRATE (86% aged 18-21 and 51% aged 22+ in spring 1999) than among those supplying a response (43% aged 18-21 and 21% aged 22+). Of course, the presence of non-response to the HRRATE question does contribute further to the selectivity of the subsample of respondents for whom the direct variable is measured.

2.3. Comparison of Data on Derived and Direct Variables

We now investigate measurement error by exploratory analysis of the differences between the derived and the direct hourly pay variables. We explore evidence for the thesis, suggested by the last two sections, that the direct variable is less prone to measurement error than the derived variable. In section 3, we shall develop a method of adjusting for measurement error, based upon this thesis.

We first consider the marginal distributions of each variable for those respondents for whom values of both variables are available. Figure 1 displays the (weighted) cumulative empirical distribution functions of each variable for the June-August 1999 quarter. The distribution of the derived variable is much more dispersed than that of the direct variable, which is indicative that the derived variable is subject to more measurement error (Chesher, 1991; Fuller, 1995). As observed also for subsequent quarters, quantiles of the distributions below the 10 percentile are lower for the derived variable and quantiles above the median are higher. The presence of positive definitional errors in the derived variable, due to additions to basic pay, may provide some explanation for the higher upper quantiles of the derived variable, but does not explain the pattern for the lower quantiles, which seems more plausibly the result of random measurement error.

[Figure 1 about here]

Figure 2 displays a scatterplot of the direct variable against the derived variable for the 4130 employees aged 22+ where both variables are observed, with lines marked at the NMW. The distribution of the direct variable shows a strong element of truncation at the NMW. There is no corresponding pattern for the derived variable, suggesting that the NMW effect is masked by measurement error. In addition, the derived variable displays many more absurdly low values, for example in summer 1999 there were 71 cases where the derived variable was less than £1 per hour. Figure 2 clearly shows that there are many discrepancies between the values of the two variables and that sometimes these discrepancies can be large.

[Figure 2 about here]

In order to investigate the thesis that the discrepancies between the two variables are mainly the result of measurement error in the derived variable, we consider the sources of definitional error in this variable listed in section 2.1. The first source is that the derived variable may include additions to basic pay, such as overtime or bonuses, unlike the direct variable. Table 2 shows the (unweighted) distribution of the discrepancies between the two variables according to whether the employee receives additions to basic pay. The proportion of respondents reporting such additions is typically substantial, almost 30% in Table 2, and it does, indeed, appear that the discrepancies tend to be greater if there are additions to basic pay.

[Table 2 about here]

The second source of definitional error in the derived variable is that its denominator refers to usual hours worked whereas the numerator refers to actual last gross pay. Thus we may expect a difference when the last gross pay was not the same as usual. The unweighted distribution of the discrepancies is summarised in Table 3 according to whether or not the last pay is reported to be the same as usual or else the respondent reports that there is no 'usual amount' to their pay. As conjectured, there are much greater discrepancies when the last gross pay is reported to differ from usual or, to a lesser extent, if it is reported that there is no usual amount.

[Table 3 here]

Cross-tabulations of these discrepancies by occupation show much higher proportions of cases with the direct variable exceeding the derived variable by large amounts for professional occupations. It seems plausible that this is a result of individuals from these occupations tending to report their usual hours as greater than their actual paid hours, leading to negative error in the derived variable.

The third source of definitional error, arising when a respondent reports that their last pay received was for a period of less than a week, occurred for less than one per cent of cases in summer 1999, but the discrepancies did indeed appear to be larger for these cases, with the derived variable exceeding the direct variable by over £2 in 25 out of 32 cases. It seems likely that in many of these cases the respondent worked for more than half a week so that the rule of doubling the actual gross pay has led to overstatement of the derived variable.

Table 4 shows the unweighted distribution of the discrepancies if all the above three sources of definitional error are excluded. Roughly 80% of the discrepancies exceeding £2 in absolute value in the 'All' columns in Tables 2 and 3 are removed by this restriction. This explanation of the discrepancies by factors known to produce errors in the derived variable further supports the thesis that the derived variable is more prone to error than the direct variable.

[Table 4 about here]

Turning to sources of reporting error, the effect of proxy reporting is considered in Table 5 (estimates are unweighted). Proxy reporting may be expected to lead to error in both the derived and direct variables. In fact, there is no evidence in Table 5 of greater discrepancies for proxy respondents. There is even a slight indication of smaller discrepancies, with proxy respondents showing greater consistency in their responses to the different questions.

[Table 5 about here]

Figure 2 displayed the very different truncation effects of the NMW for the two variables in one quarter. Further evidence of the NMW effect is obtained by examining how the proportion of people aged 22+ with pay below the NMW rate of £3.60 changed in the months before and after the introduction of the NMW in April 1999. Considering only cases where both variables were available, the proportion with the derived variable below £3.60 fell from 14.8% in March (weeks 1-4 of the March-May quarter) to 10.9% in May (weeks 9-13), whereas the proportion with the direct variable below £3.60 fell from 11.4% to 2.1%. The latter much steeper change is much more plausible and suggests again that patterns of the derived variable are masked by measurement error.

3. Imputation for Missing Values of Hourly Rate

3.1. Basic Approach and Assumptions

We conclude from the previous section that the direct variable is preferred to the derived variable as a measure of hourly pay. The main problem with the direct variable is that it is missing for a large proportion of the sample. Moreover, it is clear from section 2.2. that the direct variable is reported selectively so that considerable bias could arise if estimation was based solely on cases for which the direct variable is measured. In this section this missing data problem is addressed by an imputation approach in which missing values are replaced by imputed values. One alternative approach would be to replace missing values of the direct variable by values of the derived variable. Figure 1 suggests, however, that such an approach could still lead to appreciable upward bias in low pay proportions and the numerical impact of using this approach is illustrated later in Figure 3. Another alternative approach, using weighting, is discussed in section 4.

Let y_{1i} and y_{2i} denote the values of the derived and direct variables respectively for job i. Let s_1 denote the sample of jobs for which y_{1i} is recorded and s_2 denote the subsample of s_1 for which y_{2i} is observed. Letting r_i be the indicator variable for whether y_{2i} is observed ($r_i = 1$ if y_{2i} is observed, $r_i = 0$ if missing), we may write $s_2 = \{i \in s_1 : r_i = 1\}$. Let y_i^I denote an imputed value of hourly rate for a job where the value of the direct variable is missing ($r_i = 0$) and let $\tilde{y}_i = y_{2i}$ if r_i=1 and $\tilde{y}_i =$ y_i^I if r_i=0. Then $F_D(y)$ is estimated as in (2) by replacing y_i by $\tilde{y}_i$ (and s_D by $s_{1D} = s_1 \cap D$), that is by

$$\tilde{F}_D(y) = \sum_{s_{1D}} w_i I(\tilde{y}_i \le y) / \sum_{s_{1D}} w_i . \qquad (3)$$

The aim is to specify a method of imputation for which this estimator is approximately unbiased. It is assumed here that the weights w_i adequately compensate for any selective non-response in s_{1D} so that the only possible source of bias comes from the use of $\tilde{y}_i$ rather than the true value y_i. To assess bias, it is supposed first that the imputed values y_i^I may depend upon values y_{1i} of the derived variable as well as values x_i of covariates measured in the survey for all jobs in s_1. A model is then assumed in which, for each job i, the values y_{1i}, x_i and r_i are realised values of the random variables Y_1, X and R respectively. Likewise, the y_{2i} are treated as realisations of the random variable Y_2 if $r_i = 1$ and the y_i^I as realisations of Y^I if $r_i = 0$. We define $\tilde{Y}$ as Y_2 if R=1 and Y^I if R=0. A sufficient condition for $\tilde{F}_D(y)$ to be approximately unbiased for $F_D(y)$ is then that

$$E\left[I(\tilde{Y} \le y) | Y_1, X, R\right] = E\left[I(Y \le y) | Y_1, X, R\right] , \qquad (4)$$

where E denotes expectation under the model and membership of the set D has been subsumed in X.

Given the evidence in Section 2, it is assumed that the direct variable is measured without error and we set Y_2=Y for all jobs i so that equation (4) may be replaced by

$$E\left[I\left(\tilde{Y} \le y\right) \middle| Y_1, X, R\right] = E\left[I\left(Y_2 \le y\right) | Y_1, X, R\right] . \qquad (5)$$

If the direct variable is subject to error, then we may view $\tilde{F}_D(y)$ as an estimator of the distribution of the direct variable Y_2, assumed well-defined across all jobs in D. Whether or not the direct variable is subject to error, condition (5) defines the property desired for the imputation method. This condition holds automatically if R=1 since in this case $\tilde{Y} = Y_2$. Hence, the critical requirement for the imputation method is that (5) holds when R=0 and, if this holds for all y, we may write the condition as:

$$\left[Y^I \middle| Y_1, X, R = 0\right] = \left[Y_2 | Y_1, X, R = 0\right], \qquad (6)$$

where $\left[Y^I \middle| Y_1, X, R = 0\right]$, for example, denotes the conditional distribution of Y^I given Y_1, X and $R = 0$. Thus, we would ideally like the method to generate imputed values from the conditional distribution of Y_2 given the values of Y_1 and X and the condition that R=0. For, if this could be achieved then it follows from the above argument that $\tilde{F}_D(y)$ would be approximately unbiased for $F_D(y)$.

A basic problem with drawing imputed values Y^I from the conditional distribution of Y_2 given Y_1, X and R=0 is that, by definition, Y_2 is only observed if R=1 and hence the conditional distribution $\left[Y_2 \middle| Y_1, X, R = 0\right]$ cannot be fitted to the data directly. This is the usual identification problem with missing data modelling (Little and Rubin, 2002) and some identifying assumption is required. We make the following missing at random (MAR) assumption, common in the missing data literature (Little and Rubin, 2002).

Assumption (MAR): R is conditionally independent of Y_2 given Y_1 and X.

An alternative statement of this assumption is that the regression relationship between the hourly rate variable and the predictor variables is the same for individuals for which the hourly rate variable is measured and those for which it is missing, that is

$$\left[Y_2 \middle| Y_1, X, R = 0\right] = \left[Y_2 \middle| Y_1, X, R = 1\right]$$

Given this assumption, an imputation scheme will generate unbiased estimators if the imputed values may be drawn from the conditional distribution of Y_2 given Y_1, X and R=1. In section 3.2 we consider how to achieve this condition by fitting a regression model to the survey data $\{y_{2i}, y_{1i}, x_i; i \in s_2\}$ for which y_{2i} is observed, with y_{2i} as the dependent variable and with y_{1i} and x_i as the covariates.

A possible alternative identifying assumption to MAR is that R is conditionally independent of Y_1 given Y_2 and X. This is referred to as the *common measurement error model* assumption since it assumes that the measurement error model defined by the conditional distribution of Y_1 given Y_2 and X is the same for those reporting the direct variable and those who do not. One possible rationale for this model is that it may be more plausible for R to have a direct dependence on true pay, Y_2, than upon measured pay, Y_1 (conditional on X). Nevertheless, like MAR, this is a strong assumption, since it is plausible that the distribution of errors in reporting the components of the derived variable will differ depending on whether pay is based on an hourly rate, and it appears to be more difficult to conduct reliable inference under this assumption than the MAR assumption (see also section 5). In any case, as in standard missing data problems, it is not possible to use the observed data to test between the validity of these two assumptions since, Y_2 is unobserved when R=0. The distinction between the two assumptions may not be critical if the covariate information in the LFS, denoted here by

X, is sufficiently rich for either assumption to be a reasonable approximation. In particular, a rationale for the proposed approach based upon the MAR assumption is that, although it is likely that R will be unconditionally associated with the true pay rate, the predictive power of X in combination with Y_1 may be expected to be sufficiently strong to make the conditional association between R and the true pay rate negligible, given this information. If in fact the common measurement error model did hold and the residual conditional association were non-negligible then it might be anticipated that the conditional association between R and the true pay rate would be negative, since this is its expected sign in the absence of control for covariates. In this case, imputation based upon a model fitted to cases with R=1 would tend to under-impute the values of Y_2 for cases with R=0, leading to over-estimation of the proportion of low paid. Dickens and Manning (2002) provide a related argument that the number of low paid may be overestimated by a method based upon the MAR assumption if the common measurement error model holds and they suggest that estimates based upon the MAR assumption be viewed as upper bounds. To avoid such bias, it seems desirable to consider as rich a set of covariates, X, as possible. The effect of this choice is examined empirically in section 3.3.

In the next section we consider how to implement the imputation method based upon a regression model for Y_2 given Y_1 and X, fitted to the data $\{y_{2i}, y_{1i}, x_i; i \in s_2\}$. This model requires specification only as a conditional probability distribution for the purpose of prediction, to make the MAR assumption plausible and to improve efficiency of estimation. No assumption is made about the exogeneity of Y_1 or X.

3.2. Imputation Method

A simple approach would be to impute using the usual predicted values of Y_2 from the least squares regression on Y_1 and X. This would, however, artificially reduce the variation in the estimated distribution of interest (Little and Rubin, 2002, p.64), leading to potentially serious underestimation of proportions in the lower tail of the distribution. One way of preserving the variation in the distribution is to form the imputed values by adding randomly selected residuals to these predicted values (Little and Rubin, 2002, p.65) and this approach was explored. An alternative approach considered was a donor imputation method, using the estimated regression model to select donors by 'predictive mean matching' (Little, 1988).

Some results for these two imputation methods are given in Stuttard and Jenkins (2001, Table 2). The donor method has the advantage of being more robust to model specification and, in particular, to the implied measurement error process. This appeared to be particularly important around the NMW rate where a large spike was present in the distribution of the direct variable. The donor method preserved this feature in the imputed values, whereas the regression method with added residuals tended to smooth this spike out. Since it was of particular interest to estimate the proportion of jobs paid below the NMW rate, the donor method was chosen to avoid artificial bias.

The method of donor selection involves first determining predicted values $\hat{y}_{2i}$ from a regression model for Y_2 given Y_1 and X, fitted to the data $\{y_{2i}, y_{1i}, x_i; i \in s_2\}$. In specifying covariates, we look primarily for predictors of measurement error in hourly pay (c.f. Bound et. al.,1994, and Brownstone and Valletta,1996), since the regression models how X and Y_1 predict $\Delta = Y_2 - Y_1$. Since the predictor Y_1 is subject to error, however, we look also for direct predictors of hourly pay, as might appear in a wage equation (e.g. Machin, 1996).

The basic regression model employed here is that specified by Stuttard and Jenkins (2001) using standard model selection and diagnostic techniques:

$$\ln\left(y_{2i}\right) = \alpha + \beta \ \ln\left(y_{1i}\right) + \delta' x_i + \in_i ,$$

where the dependent variable is the logarithm of the direct variable, the covariates include both the logarithm of the derived variable as well as variables for full or part-time status, occupation, educational qualifications, length of time employed, industry, region of residence, firm size and some personal characteristics such as marital status and the specification of the distribution of the disturbance term is discussed below. Least squares estimates of the coefficients are given in Table 6. The use of models with other choices of covariates is discussed in section 3.3. Before fitting the model above, values of both y_{2i} and y_{1i} obtained from non-spouse proxy respondents were scaled to adjust for systematic measurement error, as described by Stuttard and Jenkins (2001). The replacement of this adjustment by the incorporation of proxy response status directly into the model is currently being explored.

[Table 6 here]

In a donor imputation method, the imputed value y_i^I for an employee for whom y_{2i} is missing (the 'recipient') is set equal to the value of y_{2i} of a 'donor' employee for whom y_{2i} is recorded. The original predictive mean matching method (Little, 1988) involves selecting the donor to be the 'closest ' to the recipient unit with respect to the (least squares) predicted value $\hat{y}_{2i} = \exp\left[\hat{\alpha} + \hat{\beta}\ln\left(y_{1i}\right) + \hat{\delta}' x_i\right]$. The method has been extended (e.g. Heitjan and Little, 1991) to define a set of potential donors which are close to a given recipient with respect to $\hat{y}_{2i}$ and then to select the donor from this set. This extended method was adopted here with the aim of ensuring that the random variation in the conditional distribution of Y_2 given Y_1 and X is preserved. In order to define sets of potential donors, pay rates were first divided into a series of bands with width 50p per hour, subject to there being at least ten observed cases in each band, and with the top band consisting of all cases over £15 per hour. The set of potential donors for a recipient was then defined as those cases with y_{2i} observed and values of $\hat{y}_{2i}$ falling into the same band as the recipient's value of $\hat{y}_{2i}$. The donor was then selected from this set at random. Those employees in professional and associate professional occupation groups were treated separately, because of the distinctness of the distribution of the regression residuals for these groups. The essential distributional assumption in the model is that the conditional distribution of Y_2 given $Y_1 = y_{1i}$ and $X = x_i$ depends on y_{1i} and x_i only via the 'single index' $\beta \ln(y_{1i}) + \delta' x_i$ and that the coefficients of this index are estimated consistently by least squares so that the values y_{2i} of potential donors for a recipient with $Y_1 = y_{1i}$ and $X = x_i$ are drawn from a close approximation to the conditional distribution of Y_2 given $Y_1 = y_{1i}$ and $X = x_i$. In particular, the variance of Y_2 given $Y_1 = y_{1i}$ and $X = x_i$ may depend upon $\beta \ln(y_{1i}) + \delta' x_i$.

In order to avoid donor values having disproportionate influence on the resulting estimates and to minimise the variance inflation of these estimates, donors were selected 'without replacement'. Thus, once a donor was used it could not be used again until all the potential donors within the band had been used. In addition, the imputation method was protected against outlier effects by excluding as potential donors, those cases where the residual $y_{2i} - \hat{y}_{2i}$ fell outside the 0.01 or 0.99 quantiles of the distribution of these residuals.

The stochastic nature of the imputation method introduces an additional component of variance due to imputation in the resulting estimates (Shao and Steel, 1999). For estimates based upon large subgroup sample sizes, this additional component appears to be relatively minor compared to the potential bias reduction impact. For some small domains, such as 18-21 year olds, the impact appears not to be negligible. For example, ten estimates of the proportion of those aged 22+ with pay below the NMW obtained from ten imputed datasets in which the same imputation method was repeated independently 10 times, ranged from 3.3% to 3.6%, whereas ten corresponding estimates for those aged 18-21 ranged between 2.8% and 4.7%. To address this issue the imputation method was repeated independently 10 times and the resulting estimates of $F_D(y)$ averaged across the multiply imputed datasets. This repetition has no effect on the expectation of the resulting estimate but reduces the component of variation due to random imputation. This use of multiple imputation (Rubin, 1996) is sometimes called fractional imputation.

3.3. Results

Values of the proposed estimated distribution $\tilde{F}_D(y)$ for the 22+ age group using the above donor imputation method are presented in Figure 3, with three other estimated distributions for comparison, each based upon the same weighted expression in (3) with $\tilde{y}_i$ given by (a) the derived variable, (b) the direct variable when observed and the derived variable otherwise and (c) the direct variable when observed and the regression imputation $y_i^I = \hat{y}_{2i}$ otherwise. The differences between the corresponding estimates of the proportions paid below the NMW are substantial. Using the derived variable for all cases, the estimate is 6.6%. This estimate is reduced to 4.1% if the direct variable is used instead when it is observed. A reduction is expected from Figure 1. The size of the reduction of about 40% was also found in the two subsequent quarters. The relative effect of, in addition, using imputed values when the direct variable is unobserved, is even greater, with an estimate of 1.5% for the proposed method. For this and the two subsequent quarters, the estimated proportion based upon the derived variable is four or five times higher than the estimate obtained from the proposed imputation method.

The proposed estimated distribution in Figure 3 displays a plausible 'kink' at the minimum wage of £3.60, in contrast to the distribution of the derived variable, which shows no such effect. Further evidence in support of the proposed estimated distribution is that it is closer to the distribution estimated from the NES. The estimated distribution function based upon regression imputation tends to be to the right of that for the proposed method, as expected.

[Figure 3 about here]

The robustness of the proposed method is now assessed by studying the changes in three estimated proportions for the 18+ age group under six modifications of the method, with results presented in Table 7. The magnitude of the changes in the estimates of the proportion below the NMW may be assessed relative to a standard error of about 0.15% (for the estimate of 1.53%), estimated by combining conventional LFS variance estimation methodology with a method developed by Beissel (2002) to assess the effect of imputation.

Alternative models: We noted in section 3.1 the potential importance of the choice of covariates and here consider two alternative specifications. The first is a more detailed model, which was specified to improve the fit of the model in Table 6 as far as possible, while maintaining stability of the estimated coefficients for data from successive quarters. The latter condition was imposed to avoid generating spurious quarterly changes in estimates of the pay distribution, which do not reflect genuine changes. This model selection approach led to the additional inclusion of the variables: gender, age (linear, quadratic and two youth indicator variables), temporary vs. permanent contract, ever worked overtime, pay period less than monthly, whether additions to basic pay and whether last pay same as usual together with additional indicator variables for occupation, qualifications, industry and region, a quadratic term in ln(derived variable) and an interaction term between ln(derived variable) and whether last pay same as usual, and to the exclusion of the head of household variable. A second much simpler model was also considered, excluding the head of household, married, months employed, size and region variables in Table 6 and adding gender and age (as a simple linear term). The effects of replacing the model in Table 6 by these two models and keeping all other aspects of the imputation methods the same is shown in Table 7. The more detailed model has also been studied for more recent quarters and the effects are smaller, but always in the same direction. Although the effects of changes to the model are not large, the consistent finding that the estimated proportion below the NMW (and below or at the NMW) decreases as the complexity of the model increases agrees with the theoretical direction of the effect of inadequate control for covariates under the common measurement error model, discussed in section 3.1. This suggests that the more detailed model is to be preferred, provided the resulting coefficients are not subject to sampling variation so large as to generate spurious changes in the estimated proportions over time.

Alternative Imputation Methods: Four departures from the proposed imputation method were considered. First, employees in professional and associate professional occupations were not treated separately. This effectively simplifies the assumed model by not allowing for a different distribution of the regression residuals for these occupations. A small increase in the estimated proportion at the NMW is observed. The second modification was to reduce the band width from 50p to 25p. Again this change had little effect (other specifications of the bands were also considered with little impact). Sampling donors with or without replacement also had little impact. The largest effects in Table 7 are observed when cases with outlying residuals are not excluded as potential donors. The effect is to increase the estimated proportions below and at the NMW, because more cases with very low values of the direct variable (those with large negative residuals) became eligible to act as donors. The most appropriate way to treat outliers requires further research. The proposed approach restricts the weight (see section 4) attached to cases with 'surprising' values of the direct variable and thus provides protection against the possibility that some of these values are erroneous.

4. Imputation and Weighting

An alternative to imputing for missing values of the direct variable, is to apply weights $\tilde{w}_{Di}$ to the sample $s_{2D} = s_2 \cap D$ to give the following weighted estimator of $F_D(y)$:

$$\hat{F}_{\tilde{Y}D}(y) = \sum_{s_{2D}} \tilde{w}_{Di} I(y_i \le y) / \sum_{s_{2D}} \tilde{w}_{Di} . \qquad (7)$$

The estimator of $F_D(y)$ implied by the proposed imputation approach can in fact be represented in this form, where $\tilde{w}_{Di}$ is w_i plus the sum of the weights w_j for those units j in D for which i is the donor. With multiple imputation this representation is approximate with $\tilde{w}_{Di}$ equal to w_i plus the average across multiple imputations of such weights w_j.

A more direct approach to weighting, proposed by Dickens and Manning (2002), is to take $\tilde{w}_{Di}$ as the reciprocal of the estimated probability that R=1, the *propensity score*. This score might be multiplied by the survey weight w_i to allow also for individual non-response. Under the MAR assumption, the propensity score may be estimated by fitting a regression model with r_i as the dependent variable and y_{1i} and x_i as covariates. The specification of this regression model replaces the specification of the imputation regression in section 3.2. One advantage of this approach is that it is non-stochastic and does not need to be applied repeatedly. Possible advantages of the imputation approach are that: it may be more efficient since it may make use of covariates which are predictive of Y_2 but unrelated to R, whereas propensity score weighting is 'essentially blind to efficiency concerns' (Rubin,1986); it may incorporate data modifications at the individual level, such as the proxy adjustment and outlier adjustments above and it provides imputed values for use in further analyses relating pay to other variables observed for the full sample s_1. Dickens and Manning (2002) provide some empirical comparisons of the two approaches and find they produce similar results.

5. Further Research

Further research on the properties of these estimates is being undertaken through simulation, under alternative assumptions about the data generation process. The properties of the estimation methodology under two alternative imputation methods are also being investigated. One approach is based on the original predictive mean matching method of Little (1988) using nearest neighbour imputation. Chen and Shao (2000) demonstrate theoretically the consistency of such an approach for distribution function estimation. One potential attraction of this approach is that it is not dependent upon the arbitrary choice of the width of the bands used for donor selection. Inference under the alternative common measurement error model assumption discussed in section 3.1. is also being explored using iterative procedures.

This paper has focussed on point estimation, but it is also necessary to estimate standard errors, especially for small domains, such as 18-21 year olds. Conventional standard error estimates, treating the imputed data as real, are likely to be too small. There is a growing literature on variance estimation in the presence of imputed data (e.g. Shao and Steel, 1999). Beissel (2002) has developed a standard error estimation approach for the estimators described in this paper, under the assumption that sample jobs are independent. Extensions are currently being researched to allow for the clustering of jobs within households that occurs in the LFS.

6. Conclusions

Measurement error makes it difficult to estimate low pay proportions from LFS data accurately. There is strong evidence that the directly measured hourly rate variable, introduced in 1999, is subject to less measurement error than the derived measure based on established LFS variables. The problem with this new variable is that it is missing for a large number of cases. ONS has addressed this problem by an imputation approach[1], which leads to substantially reduced estimates of low pay proportions. There is evidence from the shape of the resulting estimated distribution function around the NMW and from comparisons with the NES that the imputation approach provides improved estimates.

Ongoing research is investigating the properties of the estimated distribution under alternative assumptions and considering how these properties might be improved by development of the imputation method.

Appendix. Labour Force Survey Methodology

The LFS collects data on about 60,000 households in Great Britain per quarter. The sample is made up of five subsamples, each consisting of households living at about 12,000 addresses selected from the Postcode Address File with equal probabilities by stratified systematic sampling. The sample includes all adults living at the selected addresses. Between each quarter one of the five subsamples is 'rotated out' and replaced by a newly selected subsample. As a result, each subsample of addresses remains in the sample for five successive quarters or 'waves' of data collection. The resulting sample of adults is clustered by address but not otherwise by geography. Interviews over the five

[1] The imputation method described in section 3.2. has been applied by ONS to produce estimates from 1998 to 2001. They have also produced estimates using a 'nearest neighbour' approach (see section 5) rather than the 'band' approach. These revised estimates for 1998 to 2001 and estimates for 2002, using the nearest neighbour approach, were released by the ONS in October 2002. Details of the revisions to the methodology are on the ONS website. Any queries concerning the methodology previously used by the ONS should be addressed to Nigel Stuttard at nigel.stuttard@ons.gov.uk.

waves are held either face-to-face or by telephone. Proxy responses from other household members may be used. Information on earnings has been collected since September 1997 at the first and fifth waves for each subsample, that is from about 24,000 households per quarter, generating a sample of about 17,000 employees per quarter. Weights are constructed to compensate for differential non-response. Separate weights are constructed for earnings data using population-level information on sex, age, region, occupation, industry and whether full or part-time (Elliot,1999).

References

Beissel-Durrant, G. (2002) Variance estimation for estimates of a pay distribution based on imputed survey data. Working paper, Department of Social Statistics, University of Southampton.

Bollinger, C.R. (1998) Measurement error in the Current Population Survey: a non-parametric look. *Journal of Labor Economics*, **16**, 576-594.

Bound, J., Brown, C., Duncan, G.J. and Rodgers, W.L. (1994) Evidence on the validity of cross-sectional and longitudinal labor market data. *Journal of Labor Economics*, **12**, 345-368.

Bound, J. and Krueger, A.B. (1991) The extent of measurement error in longitudinal earnings data: do two wrongs make a right? *Journal of Labor Economics*, **9**, 1-24.

Brownstone, D. and Valletta, R.G. (1996) Modeling earnings measurement error: a multiple imputation approach. *Review of Economics and Statistics*, **78**, 705-717.

Chen, J. and Shao, J. (2000) Nearest neighbour imputation for survey data. *Journal of Official Statistics*, 16, 583-599.

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

Dickens,R. and Manning, A. (2002) Has the National Minimum Wage reduced UK wage inequality? Discussion Paper 533. Centre for Economic Performance, LSE.

Duncan, G.J. and Hill, D.H. (1985) An investigation of the extent and consequences of measurement error in labor-economic survey data. *Journal of Labor Economics*, **3**, 508-32.

Elliot, D. (1999) Report of the Task Force on Weighting and Estimation. GSS Methodology Series no. 16, Office for National Statistics.

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

Heitjan, D.F. and Little, R.J.A. (1991) Multiple imputation for the fatal accident reporting system. *Applied Statistics*, **40**, 13-29.

Little, R.J.A. (1988) Missing data adjustments in large surveys (with discussion). *Journal of Business and Economic Statistics,* **6**, 287-297.

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data*. 2nd Ed. New York: Wiley.

Machin, S. (1996) Wage inequality in the UK. *Oxford Review of Economic Policy*, **12**, 47-64.

Mellow, W. and Sider, H. (1983) Accuracy of response in labor market surveys: evidence and implications. *Journal of Labor Economics,* **1**, 331-344.

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.

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.

Shao, J. and Steel, P. (1999) Variance estimation for survey data with composite imputation and non-negligible sampling fractions. *Journal of the American Statistical Association*, **94**, 254-265.

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.

Wilkinson, D. (1998) Towards reconciliation of NES and LFS earnings data. *Labour Market Trends*, May 1998, 223-231.

Table 1. Summary Measures (Weighted) for Distribution of Derived Variable for All Cases and Only for Cases Where Direct Variable Reported, Jun-Aug 1999

Summary Measure	All Cases	Cases with Direct Variable Reported
Mean	8.19	5.57
standard deviation	5.78	3.05
1 percentile	1.63	1.67
5 percentile	3.02	2.82
25 percentile	4.63	3.85
50 percentile	6.67	4.80
75 percentile	10.00	6.50
95 percentile	17.84	10.54
99 percentile	29.49	16.00

Table 2. Distribution (%) of Discrepancies by whether Additions to Basic Pay

Range of	Addition to Basic Pay			All
discrepancies	Yes	No	Don't Know	
[£2.00, ∞)	3.1 (0.5)	3.2 (0.3)	0.0 (0.0)	3.2 (0. 2)
(£0.00, £2.00)	25.1 (1.2)	42.2 (0.9)	62.5 (12.1)	37.4 (0. 7)
[£0.00, £0.00]	2.1 (0.4)	13.4 (0.6)	12.5 (8.2)	10.2 (0.4)
(-£2.00, £0.00)	49.5 (1.4)	35.4 (0.8)	18.8 (9.7)	39.3 (0.7)
(-∞, -£2.00]	20.2 (1.1)	5.9 (0.8)	6.3 (6.0)	9.9 (0.4)
All	100% (n=1315)	100% (n=3351)	100% (n=16)	100% (n=4682)

Note: discrepancy = direct variable – derived variable; percentages are unweighted; standard errors (%), based upon binomial assumption, in parentheses

Table 3. Distribution (%) of Discrepancies by whether Last Pay Same as Usual

Range of discrepancies	Last Pay Same as Usual			All
	Yes	No	No Usual Amount	
[£2.00, ∞)	2.1 (0.3)	8.5 (1.3)	2.6 (0.6)	2.8 (0.3)
(£0.00, £2.00)	40.1 (0.9)	24.3 (2.1)	34.0 (1.7)	37.4 (0.7)
[£0.00, £0.00]	13.9 (0.6)	1.6 (0.6)	5.0 (0.8)	11.1 (0.5)
(-£2.00, £0.00)	38.7 (0.9)	35.3 (2.3)	41.1 (1.8)	38.8 (0.7)
(-∞ , -£2.00]	5.2 (0.4)	30.3 (2.2)	17.3 (1.4)	9.9 (0.5)
All	100% (n=3083)	100% (n=436)	100% (n=759)	100% (n=4278)

Table 4. Distribution (%) of Discrepancies when No Additions to Basic Pay, Last Pay Same as Usual and Pay Period not less than One Week

Range of discrepancies	%
[£2.00, ∞)	1.9 (0.3)
(£0.00, £2.00)	43.0 (1.0)
[£0.00, £0.00]	16.5 (0.7)
(-£2.00, £0.00)	35.7 (1.0)
(-∞ , -£2.00]	2.8 (0.3)
All	100% (n=2490)

Table 5. Distribution of Discrepancies by Type of Response

	Personal Response	Proxy Response		All
Range of discrepancies		Spouse/Partner	Other	
[£2.00, ∞)	3.4 (0.3)	3.2 (0.6)	2.2 (0.6)	3.2 (0.3)
(£0.00, £2.00)	37.7 (0.8)	33.5 (1.7)	43.0 (2.1)	37.6 (0.7)
[£0.00, £0.00]	8.8 (0.5)	10.1 (1.1)	17.8 (1.7)	10.1 (0.4)
(-£2.00, £0.00)	39.7 (0.8)	44.2 (1.8)	30.1 (2.0)	39.4 (0.7)
(-∞ , -£2.00]	10.4 (0.5)	9.0 (1.0)	6.9 (1.1)	9.8 (0.4)
All	100% (n=3458)	100% (n=780)	100% (n=535)	100% (n=4773)

Table 6. Estimated Coefficients of Regression Model used for Imputation

Independent Variable	Coefficient	(standard error)	Mean/ Proportion*
Intercept	0.865	(0.016)	1
ℓn (derived variable)	0.388	(0.008)	1.62
part-time	-0.047	(0.007)	43%
Occupation			
Managers and admin	0.180	(0.018)	3.2%
professional	0.477	(0.025)	1.6%
associate professional	0.230	(0.018)	3.1%
craft and related	0.103	(0.010)	14.2%
clerical and secretarial	0.060	(0.010)	11.4%
personal and protective services	0.032	(0.009)	18.7%
Head of household	0.067	(0.007)	44.8%
Married	0.049	(0.006)	53.7%
Qualifications			
degree level	0.078	(0.017)	3.6%
NVQ level 1/equiv	-0.041	(0.008)	20.3%
None	-0.068	(0.008)	21.9%
pay period less than weekly	-0.229	(0.037)	0.66%
months employed	0.0002	(0.000)	66.45
size (25+ employees at workplace)	0.052	(0.006)	60.9%
Industry			
distribution, hotels and restaurants	-0.054	(0.007)	30.6%
other services	-0.060	(0.013)	6.7%
Region : London	0.079	(0.012)	6.5%

Notes: June-August 1999 data, n= 4821 employees with complete values of all variables, dependent variable is ℓn (direct variable), $R^2 = 0.62$, * final column shows means of independent variables for this sample or % for indicator variables.

Table 7. Estimates for June-August 1999 Under Modifications of Regression Model or Imputation Method

	Estimate (%)		
	Below NMW	At NMW	Between NMW and £5/hour
Proposed Method	1.53	3.94	27.40
Modifications to Method			
More detailed model	1.31	3.80	26.37
Simpler model	1.60	3.95	27.41
Professionals not treated separately	1.51	4.06	27.09
25p bands	1.51	3.90	27.31
With replacement	1.52	3.88	27.44
Outliers not excluded	2.00	4.13	28.79

Note: "at NMW" denotes estimates between the NMW and 5p above.

Figure 1: Cumulative Distributions of the Direct and Derived Variables for Cases where both Variables are Recorded, June-August 1999.

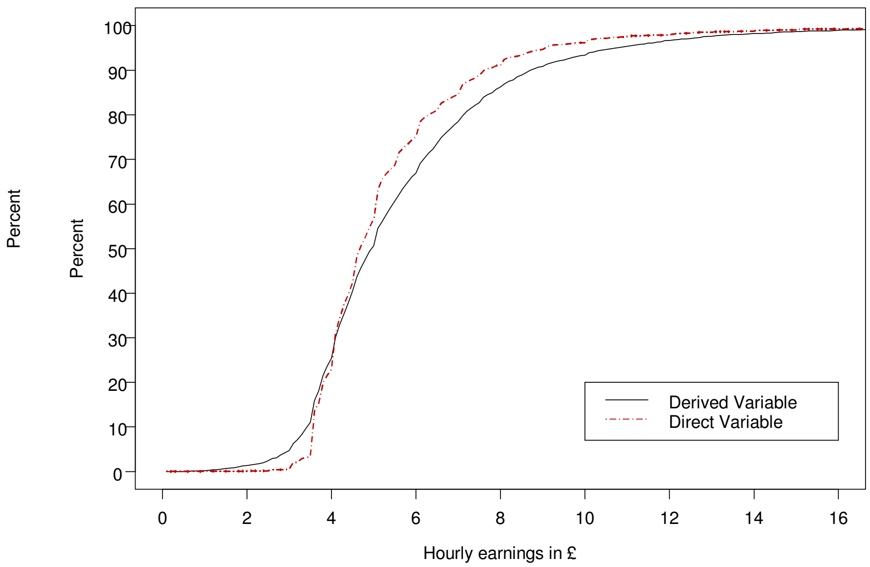


Figure 2: Scatterplot of Direct versus Derived variables with Lines Marked at NMW, for cases where both Variables are Recorded, June-August 1999.

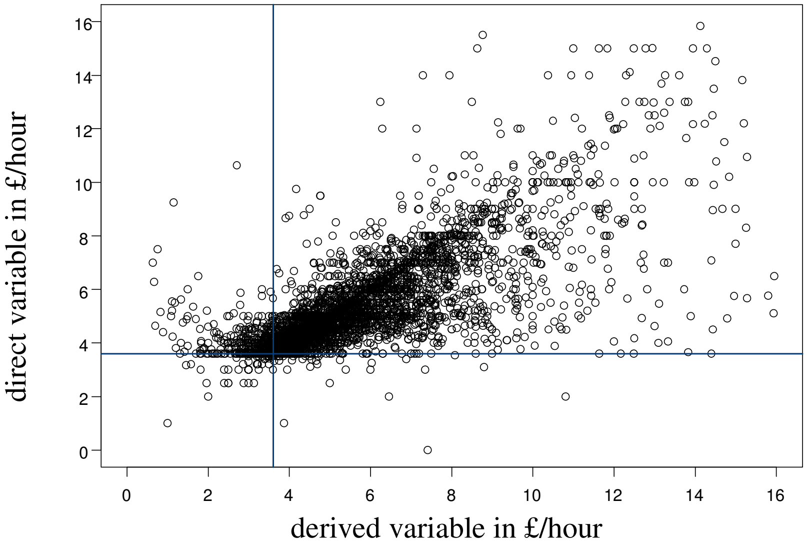
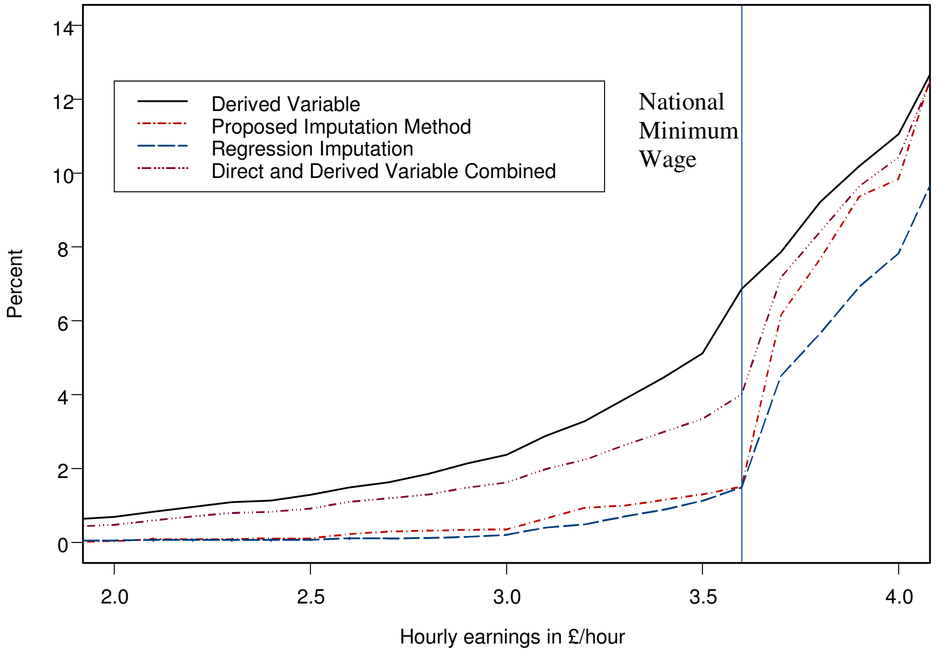


Figure 3: Cumulative distribution of hourly earnings from £2 to £4 for 22+ age group for June- August 1999

9


Derived Variable
Proposed Imputation Method
Regression Imputation
Direct and Derived Variable Combined
National
Minimum
Wage
Percent
Hourly earnings in £/hour